NEW YORK STATE TEACHER CERTIFICATION EXAMINATIONS™ (NYSTCE®)

ACADEMIC LITERACY SKILLS TEST

TEST DEVELOPMENT AND IMPLEMENTATION

DRAFT

February 2015

Table of Contents

Overview

Report on Academic Literacy Skills Test Development and Implementation

This report provides a description of test development and implementation activities for the Academic Literacy Skills Test (ALST), a new teacher certification examination that became operational in 2013 and, in 2014, became a required part of the New York State Teacher Certification Examinations (NYSTCE) program.

Below is a summary of the chapters of this report:

Chapter 1 provides an introduction to the report, including:

- a description of the New York State Teaching Standards; and
- a description of the purpose and use of the ALST.

Chapter 2 describes the test design of the ALST, including:

- a description of the ALST;
- a description of the ALST Framework;
- a description of the ALST design; and
- a description of the ALST Test Blueprint.

Chapter 3 provides descriptions of activities related to the development and validation of the ALST Framework and development of the ALST items, including:

- an overview of the ALST development activities;
- development of the ALST Framework
- validation of the ALST Framework;
- Framework review conferences;
- content validation survey;
- job analysis study;
- development of the ALST items;
- Item review conferences;
- field testing;
- marker establishment meeting;
- assembling test forms; and
- establishing performance standards.

Chapter 4 describes technical properties of the ALST scores, including:

- content-based validity evidence;

- scoring-based validity evidence;
  - scoring process
  - ALST reporting scale
    - setting the reporting scale
    - maintaining the reporting scale
- generalization-based validity evidence;
- a description of test equating;
- an overview of the scaling model;
- a report on test reliability; and
- total scaled score distribution information.

.

Chapter 5 describes score reporting of the ALST, including:

- candidate score reports;
- score reports for NYSED and New York State teacher preparation program providers; and
- an overview of Title II reporting.

Chapter 1: Introduction

The New York State Board of Regents (Regents) and the New York State Education Department (Department) are responsible for the general supervision of all educational activities in the State, including pre-kindergarten through postsecondary education, professional education and cultural education.

The Regents Reform Agenda

In an effort to achieve the Regents goals[1] for all students to have effective, high-quality teachers and leaders in every classroom and school building, the Regents developed a comprehensive reform agenda in 2009. This reform agenda seeks to strengthen teacher preparation, raise the bar for teacher certification, provide targeted support and professional development to in-service teachers, create incentives to recruit skilled teachers into high-need schools, and create school cultures that support teacher retention and encourage reflective teaching, particularly in areas with a teacher shortage. Strengthening teacher quality to improve student learning and achievement has been the driving force behind the Board of Regents and Department's policies.

To accomplish such ambitious goals, the Regents adopted a multi-pronged approach that includes the following teaching initiatives to transform teaching in New York State[2]:

- developing and adopting world class teaching and learning standards;
- implementation of new and revised exams for teacher certification, including a performance-based assessment in order to raise the bar for teacher certification;
- creating incentives for districts to implement a career ladder system for the advancement of qualified teachers in order to retain effective teachers within the profession;
- using student achievement as one of several measures to evaluate teacher performance and effectiveness in the classroom; and
- creating greater opportunities for teachers by implementing model induction and professional development programs in schools.

Teacher Certification in New York State

An individual is qualified to teach in a New York State public school if he or she is 18 years of age and possesses a New York State teacher certification.[3] The Department issues three types of

[1]Program Description Handbook 2012-2013, New York State Education Department, available at: http://www.oms.nysed.gov/budget/pro2012/toc12.html.

[2]See Board of Regents Item 1209hed (2), Dec 2009: Part II: Transforming Teaching and Learning and School Leadership in New York State, available at: http://www.regents.nysed.gov/meetings/2009Meetings/December2009/1209monthmat-new.html#hep.

[3]An individual must also subscribe to an oath of office, meet certain citizenship requirements and be fingerprinted.

certificates to individuals applying for teacher certification in the State on or after February 2, 2004: initial certificates, transitional certificates and professional certificates. Professional certificates are final, permanent certificates, and are usually the ultimate goal of those entering the profession; whereas initial and transitional certificates are time-limited, first level certificates that allow an individual to teach in a New York State classroom while building to the requirements of an initial or professional certificate, depending on the pathway the teacher chooses. By meeting specified educational and experience criteria gained during the period the certificate is valid, the initial certificate leads to a professional certificate. Candidates seeking an initial certificate must, among other requirements, complete a required course of study and pass certain required competency examinations. Candidates applying for initial certification on or after May 1, 2014 are required to pass: the Academic Literacy Skills Test (ALST), the Educating All Students Test (EAS), the edTPA (a performance-based exam),[4] and, if required by the subject the teacher desires to specialize in, a content specialty test (CST) in that particular subject. In order to obtain a transitional certificate, candidates must also pass the ALST, EAS and an appropriate CST, if required for the certificate title sought, prior to becoming eligible for the transitional certificate.

The specific requirements for each certification title are described in detail in Part 80 of the Commissioner's Regulations (8 NYCRR 80). However, regardless of how an individual chooses to become qualified to enter the teaching profession in New York State, he or she must demonstratethat the minimum knowledge, skills and abilities that teachers need have been attained before entering the classroom as the teacher of record. The requisite knowledge and skills required of teachers in New York State are described further in the New York State Teaching Standards.

The New York State Teaching Standards

The New York State Teaching Standards outline the requirements and expectations as to what teachers need to know and be able to teach effectively before and after they become certified, and as they transition through their career from novice teachers, to experienced teachers, and eventually to master teachers. For individuals who are first entering the teaching profession in New York, the Teaching Standards form the basis for assessing the minimum knowledge, skills and abilities required before a teacher enters a classroom.

Over the course of multiple Board meetings held from late 2009 through early 2010 and after reviewing data on research-based international, state and city level teaching standard models and

[4]The Board of Regents adopted an edTPA "safety net" at their April and October meetings to allow any candidate who applies for and meets the requirements of an initial certificate on or before June 30, 2015, except he/she fails the edTPA, to either: (1) take and pass the ATS-W after receipt of his/her failing score on the edTPA and prior to June 30, 2015, or (2) if the candidate had previously passed the ATS-W on or before April 30, 2014 (before the new certification examination requirements became effective) and the candidate has taken and failed the edTPA prior to June 30, 2015, the candidate will be issued an initial certificate. (This paragraph also applies to Transitional B program candidates pursuing their initial certificate).

frameworks, at its February 2010 meeting, the Board of Regents voted to develop a preliminary draft of New York State Teaching Standards to serve as the framework for the Regents Reform Agenda[5].

After researching these models the Department and the Board approved releasing the preliminary draft New York State Teaching Standards and Elements to the field for review and comment, and approved forming a Teaching Standards Work Group (Work Group). Representatives of the Work Group included, but were not limited to, educators selected in cooperation with the New York State labor unions representing teachers, principals, superintendents, faculty from teacher preparation institutions, as well as content area organizations, the National Board for Teachers, and parent-teacher groups.

In Summer 2010, the Work Group revised the preliminary draft teaching standards approved by the Board, and the Work Group's draft was released to the field for comment through an on-line survey[6]. As part of this first survey respondents were asked to comment on the clarity and appropriateness of each teaching standard and element in preparation for the Work Group to develop the performance indicators – measurable, actionable behaviors, skills and activities teachers carry out in the course of their teaching (See Appendix D). Based on the responses, which were largely positive, in August 2010, the Work Group produced a revised Standards and Elements document that was used to draft Performance Indicators[7] for each Element under the seven Standards.[8] In September 2010, the Work Group completed a first draft of the Performance Indicators, and the Regents released the New York State Teaching Standards in October 2010 to stakeholders for a final review. The Department, together with participation from the teachers' labor union, developed a second survey (see Appendix H) on the Draft

[5]International frameworks are shown in Appendix B; national and state models are shown in Appendix C.

[6]Appendices E & F provide a summary and full analysis of the survey responses.

[7]Each Teaching Standard is described by a statement which is further broken down by Elements that describe the knowledge, skills, actions, and behaviors of teachers necessary to meet that particular Standard, and the Performance Indicators describe observable and measurable actions that illustrate each Element.

[8]The Work Group also reviewed the draft NYS Teaching Standards against the revised InTASC Model Core Teaching Standards (See Appendix G) published by the Council of Chief State School Officers (CCSSO) and recently released for comment. The InTASC Standards are a set of national teaching standards detailing what teachers in the profession should be able to do and include a "focus on 21st century knowledge and skills, personalized learning for diverse learners, a collaborative professional culture, improved assessment literacy, and new leadership roles for teachers and administrators." Literacy is a prominent theme in the new InTASC Standards, and it is explicitly provided in the Standards that teachers will need to help their students build "literacy and thinking skills across the curriculum," not just in single subjects. The InTASC Standards have been adopted by many states and teacher preparation programs and are recognized by the U.S. Department of Education and national teacher education accreditation agencies. The InTASC Standards also require teachers to obtain a set of similar skills and knowledge in order to enter the teaching profession. Meeting these standards directly correlates with student success and the increase of student performance. Without proficient reading and writing skills, teachers and teacher candidates will not be able to obtain the skills and knowledge required to be effective teachers. See also InTASC Model Core Teaching Standards: A Resource for State Dialogue (2011). Available at http://www.ccsso.org/Documents/2011/InTASC_Model_Core_Teaching_Standards_2011.pdf.

Standards and released it in November 2010. This survey asked respondents whether each element of the standards was clear and understandable and, more importantly, whether each element was descriptive of what a teacher needs to know and be able to do in order to be effective in the classroom. In addition, respondents were asked whether each performance indicator was clear, understandable and measurable. Results of the survey include:

- A total of 420 respondents began the second survey, while 245 completed the entire survey. The majority of respondents were teachers, school leaders, and teacher educators, including institution faculty. The response to the second survey was overwhelmingly supportive of the second draft of the Teaching Standards.
- Over four-fifths of the respondents agreed that the Elements and Performance Indicators were clear and understandable, and approved them at a rate of 80 percent or higher. The high incidence of approval indicated clear alignment between the Teaching Standards and the knowledge and skills a teacher needs to perform his or her job responsibilities successfully.
- The overwhelming majority responded that the elements describe precisely what an effective teacher needs to know and be able to do in the classroom.
- The majority responded that most performance indicators were measurable.

The Teaching Standards Work Group reviewed the survey results and made minor revisions to the October 2010 Draft based on responses to the second survey. The Regents adopted the revised New York State Teaching Standards at its January and September 2011 meetings.[9]

As evident from the survey results, the New York State Teaching Standards directly correlate with the job of teaching and served to move the Department forward on several key Regents initiatives:

- better aligning new and existing teacher preparation programs to the Standards to ensure candidates are academically prepared to enter the classroom;
- assessing the performance and preparedness of candidates for teacher certification;
- guiding the performance evaluation of practicing teachers under their APPR; and
- identifying practice-based professional development.

At its core, the job of teaching, as clearly illustrated in the Teaching Standards, is communicating knowledge and information from the teacher (and resources made available by the teacher) to the students. The Teaching Standards are built on the principle that the knowledge, skills, and abilities detailed in the standards can only be learned, retained, and performed when an

[9]New York State Board of Regents item on Teaching Standards (December 2010). Available at http://www.regents.nysed.gov/meetings/2011Meetings/January2011/111hed3.pdf and New York State Board of Regents item on Proposed Revision to the New York State Teaching Standards (August 2011). Available at http://www.regents.nysed.gov/meetings/2011Meetings/September2011/911brca7.pdf.

individual teacher candidate and/or teacher possesses a certain level of literacy skill. The Teaching Standards require teachers and teacher candidates, as they prepare to enter the classroom, to comprehend, analyze, and synthesize a set of knowledge that demonstrates their understanding, and then validate that understanding by presenting and dissimenating that knowledge using a variety of skills and techniques necessary to ensure that students are learning

New York State's foundational belief that all teachers must acquire a minimum level of literacy knowledge, skills and abilities to effectively facilitate student learning and achievement and meet the Teaching Standards is based on multiple empirical studies (e.g., Auguste, Kihn, &Miller, 2010; Walsh & Tracy, 2004). It is essential that New York State's education stakeholders be assured that New York's teachers possess literacy knowledge, skills, and abilities before being certified for practice. In New York, literacy skills are interwoven throughout the New York State Teaching Standards, which, in turn, form the foundation for the success of teacher candidates.

Embedded within the performance indicators of the NYS Teaching Standards are literacy skills because research indicates that literacy is among the most important indicators of student success in the classroom (Auguste et al., 2010). Literacy, which is often narrowly understood to be an individual's ability to read, in its broadest but truest sense, is a measure of a person's world knowledge (Walsh & Tracy, 2004). It involves using reading, writing, speaking, listening, and viewing to gain more knowledge. In other words, the more someone is familiar with a broad range of subjects, the more literate a person is. Research has also shown that a teacher's level of literacy, as measured by vocabulary, comprehension and a deep understanding of the subject matter as measured by various types of assessments, affects student achievement more than any other measurable teacher attribute, including certification status, experience, and the amount of professional development that a teacher receives (Walsh & Tracy, 2004; Wayne & Youngs, 2003; Rice, 2003) . For example, a study of National Board Certified Teachers conducted in North Carolina found that the teacher attribute that most consistently distinguished National Board Certified Teachers from others was their level of literacy (Goldhaber, Perry, & Anthony, 2004).

The New York State Teaching Standards: The Connection to Literacy

The New York State Teaching Standards outline the skills and abilities teacher candidates need to be minimally competent in before entering the classroom as the teacher of record, and they serve as an instruction manual for in-service teachers to continuously refresh and improve their knowledge and skills through professional development. Each of the seven New York State Teaching Standards are listed below, followed by the knowledge and skills outlined in each standard. Under each standard are elements/performance indicators where the literacy skills, and abilities are needed by the teacher to carry out the requirements of that particular standard, explicitly illustrating the connection between the Standards and literacy skills (see Appendix A).

Standard 1: Knowledge of students and student learning: Teachers are required to acquire knowledge of each student, and demonstrate knowledge of student development and learning.

- Literacy abilities and skills, such as making logical inferences, drawing conclusions, analyzing the central ideas of concepts, and recognizing key supporting details are necessary to the demonstration of: (1) knowledge of child and adolescent development; (2) knowledge of current research in learning and language acquisition theories and processes; (3) knowledge of and responsiveness to diverse learning needs, strengths, interests, and experiences of all students; (4) knowledge of individual students; (5) knowledge of and responsiveness of the various economic, social, cultural, linguistic and other factors that influence student learning; and (6) knowledge of understanding of technological and information literacy and how they affect student learning[10].

Standard 2: Knowledge of Content and Instructional Planning: Teachers know the content they are responsible for teaching, and plan instruction that ensures growth and achievement for all students.

- Literacy abilities, such as determining what a text says explicitly, the author's opinion, and how and why events and ideas develop and interact over the course of a text, are essential to: (1) demonstrating content knowledge; (2) demonstrating how to connect concepts and how to engage learners; (3) demonstrating the use of a broad range of instructional strategies; (4) establishing goals and expectations; (5) designing relevant instruction; and (6) evaluating appropriate curricular materials and resources.

Standard 3: Instructional Practice: Teachers implement instruction that engages and challenges all students to meet or exceed the learning standards.

- Literacy abilities and skills such as the delineation and evaluation of arguments, the evaluation of the validity of reasoning, and the introduction of precise, knowledgeable claims, are critical to: (1) providing developmentally appropriate and standards-driven instruction; (2) communicating clearly and accurately; (3) setting high expectations and creating challenging learning experiences; (4) exploring and using a variety of instructional approaches, resources, and technologies; (5) engaging students in the development of multidisciplinary skills; and (6) monitoring and assessing student progress, seeking and providing feedback, and adapting instruction.

Standard 4: Learning Environment: Teachers work with all students to create a dynamic learning environment that supports achievement and growth.

[10]The elements/performance indicators referencing literacy skills and abilities have been paraphrased to shorten the descriptions for purposes of this report. The full list of elements/performance indicators can be found in Appendix A.

- Literacy abilities and skills, such as the use of valid reasoning to support a claim, anticipating and addressing a possible counterclaim, and choosing relevant and sufficient evidence to support claims, are important to (1) creating a mutually respectful and safe learning environment; (2) intellectually challenging and stimulating learning environment; (3) managing the learning environment; and (4) organizing and utilizing available resources.

Standard 5: Assessment for Student Learning: Teachers use multiple measures to assess and document student growth, evaluate instructional effectiveness, and modify instruction.

- Literacy abilities and skills, such as producing a conclusion, making logical inferences, choosing precise language for clarity and determining the themes of a text, are essential to: (1) designing, selecting, and using assessment tools; (2) understanding, analyzing, interpreting and using assessment data; (3) communicating information about various components of the assessment; (4) reflecting upon and evaluating the effectiveness of comprehensive assessment systems; and (5) preparing students to understand the format and directions of assessments.

Standard 6: Professional Responsibilities and Collaboration: Teachers demonstrate professional responsibility and engage relevant stakeholders to maximize student growth, development, and learning.

- Literacy abilities and skills, such as choosing precise language for clarity, establishing and using an appropriate style and tone, analyzing how specific words and sentences shape the meaning of the text, and drawing conclusions, are critical to: (1) upholding professional teaching standards; (2) collaborating with colleagues; (3) collaborating with families; (4) managing and performing non-instructional duties; and (5) understanding and complying with relevant laws and policies.

Standard 7: Professional Growth: Teachers set informed goals and strive for continuous professional growth.

- Literacy abilities and skills, such as drawing conclusions, using valid reasoning, using correct standard English grammar, usage, and spelling, and choosing relevant and sufficient evidence to support claims, are important to: (1) reflecting on practices to improve instructional effectiveness; (2) setting goals to engage in ongoing professional development; (3) communicating and collaborating with others; and (4) remaining current in their knowledge of content and pedagogy.

Given the long-standing discussion of the relationship between general teacher aptitude and student achievement, it is important to consider how particular measures of teachers' abilities

relate to student learning. Research indicates that teachers' scores on tests of verbal skills (such as vocabulary or word tests) are related to the achievement of their students (Ehrenberg & Brewer, 1995; Hanushek, 1992; Rice, 2003; Wayne & Youngs, 2003).

It is apparent through the New York State Teaching Standards mentioned above, that literacy skills (e.g., reading and writing) are an integral part of becoming a teacher, and are prerequisites for teacher candidates to obtain the skills and knowledge listed in each of the elements of the Standards. Simply put, a teacher will not be able to complete the daily task of effectively educating children without proficient reading and writing skills.

Literacy skills and abilities are fundamental for a teacher to accomplish his/her teaching and other duties effectively each and every day. Studies have shown that a teacher's academic ability, measured accurately by examining literacy skills, is directly and positively linked to student performance (Ahn & Choi, 2004; Tchoshanov, 2011). Studies have also shown that successful, effective teachers with above average student achievement gains are those who can model desired behaviors, both academically and behaviorally. Since the creation of the Common Core State Standards (CCSS) and New York's Common Core Learning Standards (CCLS), P-12 students are being held to a higher academic standard. These Standards are the result of multiple studies[11],[12],[13] focused on determining the knowledge and skills P-12 students will need in order to become college and career ready and literacy skills are inextricably intertwined in the new CCLS required of students. As the standards for P-12 students rise, teachers, not surprisingly, and as part of the Reform Agenda, are being held accountable for helping to raise student achievement to this same and demanding bar. Without proficient reading and writing skills, teachers will not be prepared to educate the children entrusted to them by our society today.

In addition, the CCSS and CCLS place stronger emphasis on students' abilities to read and write across the curriculum, which echoes the NYS Teaching Standards' and the InTASC Model Core Teaching Standards emphasis on the importance of "teachers build[ing] literacy and thinking skills across the curriculum [and] help[ing] learners address multiple perspectives in exploring ideas and solving problems."[14] New York State must assure that certified teachers in every subject area have the skills, abilities, and knowledge to integrate teaching literacy skills as part of their curriculum.

---

[11] Common Core State Standards Initiative. Common core state standards for English language arts and literacy in history/social studies, science, and technical subject Appendix A. Available at http://www.corestandards.org/assets/Appendix_A.pdf.

[12] Common Core State Standards Initiative. Common core state standards for English language arts and literacy in history/social studies, science, and technical subject Appendix B. Available at http://www.corestandards.org/assets/Appendix_B.pdf

[13] Common Core State Standards Initiative. Common core state standards for English language arts and literacy in history/social studies, science, and technical subject Appendix C. Available at http://www.corestandards.org/assets/Appendix_C.pdf

[14] InTASC Model Core Teaching Standards: A Resource for State Dialogue. *supra note 7.*

What is the purpose of the Academic Literacy Skills Test (ALST)?

As stated earlier, the ALST is one of several required certification examinations in New York, which are used to measure whether a teacher candidate has the minimum knowledge, skills and abilities to enter a classroom as the teacher of record. The Academic Literacy Skills Test (ALST) measures a teacher candidate's literacy skills (reading and writing skills) implicit in the New York State Teaching Standards and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning.

For whom is the ALST intended?

The intended population of test takers comprises all teacher candidates and/or out-of-state certificate holders who apply for an initial certificate to teach in this State on or after May 1, 2014 and teacher candidates who apply for certain transitional certificates on or after May 1, 2014.

How is the ALST used?

A teacher candidate seeking to obtain an initial certificate to teach in New York State must complete an educational program and must take and pass certain competency examinations, including the ALST. The main use of the ALST score is to determine whether a teacher candidate possesses the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning.

If a teacher certification candidate does not successfully pass the ALST, the candidate can retake the ALST an unlimited number of times; however, until the candidate successfully passes the ALST and meets all other certification requirements, including other certification examinations requirements, the candidate cannot be issued an initial certificate, nor become the teacher of record in a classroom.

Chapter 2: ALST Test Design

Description of the ALST

The Academic Literacy Skills Test (ALST) is a criterion-referenced, competency-based, computer-delivered test designed to measure a teaching candidate's literacy skills (reading and writing skills), implicit in the New York State Teaching Standards and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning. These reading and writing skills are further delineated in the New York State P-12 Common Core Learning Standards for English Language Arts and Literacy (NY P-12 CCLS for ELA and Literacy) adopted by New York State in 2010 and identified by practicing New York State educators as important or extremely important to performing various critical job tasks of an educator in the New York State.

The ALST is composed of two competencies: *Reading* and *Writing to Sources*. The content of the *Reading* competency is measured by the selected-response items (SRIs) designed to assess candidates' competence in understanding and analyzing a variety of informational and literary texts representative of those required in college courses. The content of the *Writing to Sources* competency is measured by the constructed-response items (CRIs) designed to assess candidates' ability to develop a clear, coherent, and cohesive argument to support a claim in a cogent synthesis and thorough analysis of information presented in multiple informational texts. Candidates taking the ALST receive performance information on each of the two competencies, in addition to the total scaled score. Details on the ALST reporting scale and its maintenance across test administrations can be found in Chapter 4 "Technical Properties of the ALST scores". Details on the scoring and reporting of the ALST can be found in Chapter 5 "ALST Score Reporting". The construct measured by the ALST – academic literacy skills – is described in detail in the ALST Framework.

Description of the ALST Framework

The ALST Framework defines the content covered on the test (refer to Figure 1 and Appendix I). As stated in The Standards for Educational and Psychological Testing, "the delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests" (AERA, APA, & NCME, 1999, p. 37). During the framework development stage, a careful and systematic analysis of the New York State Teaching Standards indicated the importance of reading and writing skills for the job of an educator and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning. Subsequent reviews of the framework by practicing New York State educators and educator preparation faculty provided further evidence that the competencies included in the ALST Framework are important to performing various critical job tasks of an educator in New York State (see Chapter 3 for the detailed description of the Framework Review Conferences, Content Validation Survey,

and the Job Analysis study). The contents of the resulting ALST Framework are described below.

The first part of the framework is a high-level description of the New York State educator who has the academic literacy skills necessary to teach effectively in New York State public schools. The structure of the ALST Framework provides an organizational scheme used for both test development and reporting purposes. The body of the ALST Framework includes competencies, performance expectations, and performance indicators, which are described next.

Competencies. The ALST Framework is composed of two competencies: *0001 Reading* and *0002 Writing to Sources*. The two competencies are specific areas of content on which candidates will be assessed.

Performance Expectations. The competencies are elaborated by performance expectations, which are broad, performance-based descriptions of the academic literacy skills a New York State educator is expected to have in order to teach effectively in New York State public schools.

Performance Indicators. The performance indicators provide further details about the nature and range of content covered by the competencies. They are intended to suggest the type of content that may be included in the test items associated with each competency.

Purposes of the ALST Framework

The ALST Framework is designed to communicate with audiences (e.g., candidates, educator preparation faculty) interested in the testing program. The Framework is intended to:

- facilitate preparation by educator preparation institutions;
- facilitate preparation by individual candidates;
- assist in the interpretation of test results by individuals, institutions, and the NYSED; and
- guide the development of the ALST materials.

Moreover, the purpose of the ALST Framework extends to the following:

- provide a structure for the content covered on the ALST;
- establish a link between ALST content and the Common Core Learning Standards;
- provide a structure for score reporting and score interpretation for candidates;
- guide item development; and
- support technical aspects of the ALST.

NEW YORK STATE TEACHER CERTIFICATION EXAMINATIONS™

FIELD 202: ACADEMIC LITERACY SKILLS TEST (ALST)
TEST FRAMEWORK

Reading
Writing to Sources

The New York State educator has the academic literacy skills necessary to teach effectively in New York State public schools. The teacher is capable of proficient, close, and critical reading that reflects wide, deep, and thoughtful engagement with a range of high-quality, complex informational and literary texts. The teacher demonstrates command of evidence found in texts and uses cogent reasoning to analyze and synthesize ideas. The teacher produces complex and nuanced writing by choosing words, information, and structure deliberately for a given task, purpose, and audience.

COMPETENCY 0001—READING

Performance Expectations

The New York State educator reads complex informational and narrative texts and demonstrates command of key ideas and details in the texts. The teacher determines what a text says explicitly and consistently makes logical inferences and draws conclusions based on evidence found in the text. The teacher correctly determines the central ideas or themes of a text and analyzes their development. The teacher recognizes accurate summaries of key supporting details and ideas. The teacher accurately determines an author's attitude, opinion, or point of view. The teacher analyzes how and why individuals, events, and ideas develop and interact over the course of a text.

The New York State educator demonstrates command of craft and structure in reading. The teacher accurately interprets words and phrases as they are used in a text, including determining technical, connotative, and figurative meanings, and thoroughly analyzes how specific word choices shape meaning and tone. The teacher thoroughly analyzes the structure of texts, including how specific sentences, paragraphs, and larger portions of the text relate to each other and the whole. The teacher accurately assesses how point of view and purpose shape the content and style of a text.

Performance Indicators

a. determines what a text says explicitly
b. makes logical inferences based on textual evidence

Figure 1. Excerpt from the ALST Framework

Description of the ALST Design

Item Types and Weights. The ALST is designed to measure the two competencies defined in the framework: *0001 Reading* (composed of 40 selected-response items) and *0002 Writing to Sources* (composed of two focused-response and one extended-response items). The following item types are represented on the ALST: selected-response items, focused-response items, and extended-response items. Selected-response items are scored dichotomously (1 point for a correct response and 0 points for an incorrect response). Focused and extended-response items are scored polytomously with scoring rubrics (score range from 0 to 4). To determine the total test score on the ALST test, weights are applied to each type of item. Two focused-response items each contribute 15% to the total score; the one extended-response item contributes 30% to the total score; and all scorable selected-response items contribute 40% to the total score. See Table 1 for the ALST test design.

The weights of item types and the test score associated with each item type were determined by NYSED in consultation with experienced Pearson testing specialists, and were part of the context for reviews by New York State educators and educator preparation faculty during bias and content reviews of test materials and other test development meetings. Decisions regarding weights were primarily informed by content and psychometric considerations (i.e., weights that would enable reliable assessment of the range of academic literacy skills critical for performing the job of an educator).

Table 1. ALST Design

Competency	**Selected-Response**		**Constructed-Response**	
	Approximate Number of Items	**Approximate Percentage of Test Score**	**Number of Items**	**Approximate Percentage of Test Score**
0001 Reading	32 scorable 8 non-scorable	40%	--	--
0002 Writing to Sources	--	--	2 focused-response 1 extended-response	15% each 30%
Total	**40**	**40%**	**3**	**60%**

Testing time. NYSED's decisions regarding the amount of time to allocate to each component of the ALST were informed by Pearson's experience with similar testing programs and by studies completed during ALST field testing of the amount of time examinees took to complete each

component of the ALST. The total testing time allocated to candidates taking the ALS is 210 minutes. The selected-response items are designed with the expectation of total response time up to 110 minutes. Each focused constructed-response item is designed with the expectation of a response up to 20 minutes. The extended constructed-response item is designed with the expectation of a response up to 60 minutes. Candidates are free to set their own pace during test administration.

Selected-response items. The selected-response items included on the ALST provide observations of candidate ability across the required range of skills. The selected-response items require candidates to demonstrate that they are capable of closely reading and thoughtfully analyzing a wide range of high-quality literary and informational texts.

Each ALST test form contains a total of 40 selected-response items. Out of the 40 selected-response items, 32 of them (80%) are items that have been previously field tested (see Chapter 3) and are operational items (i.e., "scorable"). They are all scored as 0 or 1 and count towards a candidate's overall test score. The remaining 8 selected-response items (20%) are field test items that are "non-scorable" – that is, they do not count towards the total test score Embedding the non-scorable selected-response items on the operational test forms is a widely-used industry-accepted method that allows for collecting data on the psychometric characteristics of the items prior to their use as operational items.

Constructed-response items. Each ALST test form contains a set of three constructed-response items all linked to a three-part stimulus. Examinees must submit two focused responses and one extended response. Performance expectations and suggested response lengths, specific to each item type, are provided. The sets of constructed-response items are designed to be comparable in terms of difficulty across forms (see Chapter 4 for details).

Description of the ALST Test Blueprint

The ALST test blueprint is based on the ALST test design described above. The major features of the ALST test blueprint are outlined below and summarized in Table 2:

- Competency *0001 Reading* requires candidates to read five passages (including both informational and literary passages) and answer eight selected-response items for each passage, for a total of forty selected-response items. One of the five passages has a corresponding set of eight non-scorable items included for field-testing purposes. All reading passages are previously published texts or excerpts from previously published texts.

- Competency *0002 Writing to Sources* requires candidates to read a stimulus consisting of two passages representing different perspectives on a common topic, and one graphic representation of information relevant to the common topic. Candidates then complete

two focused constructed-response items (CRIs) and one extended constructed-response item (CRI). One focused CRI requires analysis and comparison of two authors' positions on a common topic and the other focused CRI requires integrating information presented graphically with information in the text. The extended CRI requires developing and supporting an argument on the same issue using all three stimuli as sources in an extended-length response.

Table 2. ALST Test Blueprint

Competency	Stimuli	Items per Cluster			Number of Items
		SRI	CRI Focused Response	CRI Extended Response	
0001 Reading	• 4 operational passages • 1 field test passage	8	-	-	• 32 operational items • 8 embedded field test items
0002 Writing to Sources	• 2 passages • 1 graphic representation of information	-	2	1	• 2 Focused Response CRIs • 1 Extended Response CRI

Chapter 3: ALST Test Development

Overview

The Academic Literacy Skills Test (ALST) was developed in accordance with the industry's best practices for licensing and employment testing, as specified in the AERA/APA/NCME Standards for Educational and Psychological Testing (1999). New York State educators and educator preparation faculty were actively involved in all stages of the ALST development (see Appendix K for the description of the characteristics of the ALST committee members). They provided critical input at various points in the test development process and ensured that all ALST materials are accurate, job-related, and free from bias. Throughout the development process, the focus remained on ensuring that the ALST test scores accurately and representatively measure a teaching candidate's literacy skills (reading and writing skills) implicit in the New York State Teaching Standards and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning. To this end, the following development activities were conducted:

- Development of the ALST Framework
- Validation of the ALST Framework
- Framework Review Conferences
- Job Analysis Study
- Development of the ALST Items
- Assessment Specifications
- Item Review Conferences
- Field Testing
- Marker Establishment Meeting
- Assemble Test Forms
- Establish Performance Standards

Development of the ALST Framework: Overview

The ALST Framework defines the content covered on the test (refer to Figure 1 in Chapter 2 and Appendix I). The structure of the ALST Framework provides an organizational scheme used for both test development and reporting purposes. NYSED and Pearson took a series of integrated steps to prepare the ALST Framework. As the AERA/APA/NCME Standards for Educational and Psychological Testing (1999) state:

> "The first step [in test development] is to extend the original statement of purpose(s), and the construct or content domain being considered, into a set of test objectives for the test that describes the extent of the domain, or the scope of the construct to be measured. The test framework, therefore, delineates the aspects (e.g., content, skills, process, and

> diagnostic features) of the construct or domain to be measured…. The delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests. The test framework serves as a guide to subsequent test evaluation." (p. 37)

The ALST test development process started with the development of the ALST Framework. As stated in The Standards for Educational and Psychological Testing, "the delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests" (AERA, APA, & NCME, 1999, p. 37).In accordance with these standards, careful steps were taken in the early stages of the ALST Framework development to ensure that the competencies outlined in the framework are appropriate targets of measurement for the ALST. First, a careful analysis of the job-related reading and writing skills described in the New York State Teaching Standards (Appendix A) was conducted (see Chapter 1 for the detailed description). Second, the reading and writing skills expected of New York State students (and thus, teachers) – as documented in the New York State P-12 Common Core Learning Standards for English Language Arts and Literacy (NY P-12 CCLS for ELA and Literacy; Appendix L) were carefully analyzed by NYSED and Pearson. Third, the ALST Framework (Appendix I) was drafted by Pearson testing experts (as described in the sections below) in close collaboration with NYSED based on these analyses of the job-related tasks of an educator in New York State. Fourth, correspondence was established between the contents of the ALST Framework and job-related reading and writing skills operationalized in the NY P-12 CCLS for ELA and Literacy. This correspondence was formally documented in the Content Correlation Table (Appendix N) and reviewed by New York State educators and educator preparation faculty. As a fifth and final step in the framework development process, the draft ALST Framework was reviewed for relevance and job-relatedness by the New York State Education Department (NYSED)-designated experts (Appendix J). Following the completion of the Framework development process, a multi-step process for collecting validity evidence for the Framework was conducted through the Job Analysis (JA) study which further confirmed the criticality of the competencies assessed by the ALST to the tasks of an educator (see detailed description of the JA Study on p. 39 and in the Appendix M).

Analysis of the Relevant Standards

Careful analysis of the relevant standards marked the beginning of the ALST Framework development process. The literacy skills (reading and writing skills) implicit in the New York State Teaching Standards (Appendix A) and reflecting the minimum knowledge skills and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning, served as the foundation for the initial draft of the ALST competencies. These reading and writing skills are further delineated in the New York State P-12 Common Core Learning Standards for English Language Arts and Literacy (NY P-12 CCLS for ELA and Literacy; Appendix L) adopted by Board of Regents in 2010.

Draft ALST Framework

Working in close collaboration with NYSED, Pearson testing experts (a team of full-time senior test development staff that included literacy experts, former educators, psychometricians, and leaders with decades of experience developing assessment materials for educator credentialing in several states across the country) drafted the preliminary ALST Framework. Analysis of the job-related academic literacy skills described in the New York State Teaching Standards (Appendix A) and documented in the NY P-12 CCLS for ELA and Literacy (Appendix L) was instrumental during this stage. After the draft ALST Framework was created, it underwent various reviews and content alignment activities, which are described next.

Content Correlation Table

The surveys conducted during the development of the Teaching Standards (see Chapter 1 for the detailed description) and the subsequent Job Analysis study of the ALST (see section on the New York State Job Analysis on p. 39 of this Chapter) made important contributions to the content validity evidence for the ALST. In addition, the content of the ALST was formally aligned with the Common Core Standards for ELA and Literacy because these are the standards to which NYS students are being held accountable.

Given that the reading and writing skills outlined in the ALST Framework were operationalized using the NY P-12 CCLS for ELA and Literacy (Appendix L), formally establishing a correspondence between these standards and the contents of the ALST Framework was the next logical step. This correspondence is reflected in the Content Correlation Table (Appendix N), which maps each one of the ALST performance indicators to the corresponding NY P-12 CCLS standard, with some indicators addressing more than one standard.

The following question guided the process of deriving the ALST Content Correlation Table:

"Is the content of the performance indicator addressed, in whole or in part, by content included in the NY P-12 CCLS for ELA and Literacy?"

Pearson testing experts (including the team of full-time senior test development staff previously described) worked in close collaboration with NYSED to develop the ALST Content Correlation Table, which can be found in Appendix N.

Review of the Draft ALST Framework

The draft ALST Framework was then submitted for a preliminary review for relevance and job-relatedness by the New York State Education Department (NYSED)-designated experts (Appendix J).

Validation of the ALST Framework: Overview

After the ALST Framework had undergone initial development, further steps were taken in order to gather additional content-based validity evidence. First, the ALST Framework was reviewed by New York State educators and educator preparation faculty (Appendix K) during the Framework Review Conference. During this effort, a Bias Review Committee (BRC) reviewed the content of the ALST Test Framework and was charged with bias prevention. In addition, the BRC also reviewed the ALST Framework for appropriateness and job-relatedness. Following the Bias Review, a Content Advisory Committee (CAC) reviewed the framework for content appropriateness and job-relatedness, incorporated bias-related comments from the BRC, and made final recommendations to NYSED regarding the ALST Framework. As a second step, Content validation (CV) surveys were conducted with New York State educators and faculty in order to gather additional evidence confirming that the content of the ALST Framework is appropriate and job-related. A full report containing CV Survey Results for the ALST can be found in Appendix O.

The third and final step in the process of collecting content-based validity evidence for the ALST Framework – conducting the Job Analysis Study – was essential for ensuring that the tasks identified by the New York State educators as critical for performing the job of an educator were covered in the ALST Framework, thereby rendering ALST competencies and performance indicators as appropriate targets of measurement for the ALST. Empirical evidence garnered through the multi-step Job Analysis Study supported the logical chain from critical job-related tasks performed by the NYS educators, to the knowledge, skills, abilities, and other characteristics (KSAOs) (referred to as "competencies" in the ALST Framework) required for performing those tasks, to the inclusion of these critical competencies in the ALST frameworks. A full report can be found in Appendix M.

Conduct Framework Review Conferences

Bias Review Committee (BRC)

Function. The main function of the ALST BRC was to ensure that all test materials are free from bias and are fair and equitable for all candidates. The BRC was charged with bias prevention, which is defined as (1) excluding language or content that might disadvantage or offend an examinee because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background and (2) including content that reflects the diversity of New York State. Prior to starting its reviews, the BRC went through an orientation during which the members were trained on the best practices in preventing bias in tests and received a copy of *Fairness and Diversity in Tests* (2009). The BRC was convened multiple times during the framework and item review stages of the test development process.

Recruitment and Selection. The ALST BRC included practicing New York State public school educators who hold permanent or professional certification in New York State and New York

State educator preparation faculty (including education faculty and arts and sciences faculty) who are preparing candidates for educator certification. Special emphasis was also placed on recruiting BRC members representative of the regional, gender, and ethnic diversity of New York State.

Several targeted recruitment efforts were conducted by the NYSED and Pearson in order to recruit and select qualified public school educators and educator preparation faculty to serve on the BRC. Applications for committee membership were elicited from public school administrators, public school teachers, state professional organizations (e.g., New York State United Teachers), and other sources identified by the NYSED. NYSED reviewed the applications and selected individuals to serve on the BRC based on a review of their qualifications (e.g., educational position, professional activities) and representation of the regional, gender, and ethnic diversity of the New York State.

Committee applicants who met the committee guidelines and indicated a willingness to serve on the BRC were considered for service based on several factors. These included:

- experience working with diverse populations and/or traditionally marginalized communities
- background/training in social justice/critical theory
- expressed sensitivity and concern for equity issues, including those affecting non-traditional communities, populations, and orientations.

Applicants who possessed the above characteristics were selected to serve on the BRC based on traditional concerns (geographic, ethnic, cultural, and gender representation). *Composition*. The ALST BRC included permanently certified public school educators and college and university faculty, including those teaching undergraduate or graduate education or arts and sciences courses in which prospective educators were enrolled. BRC members were recruited to include representation of the regional, gender, and ethnic diversity of New York State. In addition, special consideration was given to the following criteria:

- representation from professional associations and other organizations;
- representation from diverse racial, ethnic, and cultural groups;
- representation from females and males;
- geographic representation; and
- representation from diverse school settings (e.g., urban areas, rural areas, large schools, small schools).

The ALST BRC is composed of 24 members including both certified and practicing New York State educators and educator preparation faculty. Each time a meeting of the BRC is convened,

members of the Committee are invited to participate in the review of NYSTCE test materials. Refer to Appendix K for a description of the characteristics of the BRC members.

Bias Review of the ALST Framework

The ALST Bias Review Committee (BRC) systematically reviewed the draft ALST Framework. This review was conducted to help ensure that the ALST Framework was free from bias. The comments and recommendations of the BRC were then communicated to the Content Advisory Committees (CAC) during the CAC review of the draft ALST Framework.

The NYSTCE BRC meeting at which the ALST Framework document was reviewed was conducted at the Pearson office in Malta, New York, on April 16, 2012. A total of four members (including both New York State educators and educator preparation faculty, as described previously in this chapter) participated in the review of the ALST Framework.

Introduction, orientation, and training. Staff from NYSED and Pearson conducted an orientation for the bias review session. During the orientation, BRC members received information on the background and purpose of the ALST, the purpose of the framework review activity, and the review process for the ALST Framework. Prior to starting their review, committee members were trained on best practices in detecting and preventing bias in tests and received a copy of *Fairness and Diversity in Tests* (2009). The training was conducted by Pearson testing specialists with many years of experience providing similar training and facilitation for several other bias review committees engaged in the review of assessment materials for educator credentialing programs across the country.

The BRC workgroup was assigned a Pearson facilitator, who oriented committee members to their tasks, explained procedures and materials, addressed testing issues, facilitated discussions, and kept the master copy of the workgroup's recommendations. All efforts were made to ensure that each committee member was able to express his or her opinion in the review and that consensus regarding the test materials was being reached.

Materials. Each BRC member registered upon arrival by signing a Sign-in Sheet and a Confidentiality Agreement. BRC members received the following materials:

1. Orientation Manual
2. Framework Review Booklet
3. *Fairness and Diversity in Tests*

Bias Review Guidelines for Test Design and Framework. BRC members were instructed to ask themselves a set of organized questions focused on bias-related issues to review the proposed frameworks (including test competencies, performance expectations, and performance indicators).

Content Does any element of the framework, including competencies, performance expectations, and performance indicators, contain content that disadvantages a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Language Does the language used to describe any element of the framework, including competencies, performance expectations, and performance indicators, disadvantage a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Offense Is any element of the framework, including competencies, performance expectations, and performance indicators, presented in such a way as to offend a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Stereotypes Does any element of the framework, including competencies, performance expectations, and performance indicators, contain language or content that reflects a stereotypical view of a group based on gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Fairness Taken as a whole, is the framework fair to all individuals regardless of race, gender, cultural background, or other personal characteristics?

Diversity Does the framework permit appropriate inclusion of content that reflects the diversity of the New York State population?

Review procedures. BRC members reviewed the draft ALST Framework to verify that the framework meets the bias review criteria. The procedures for the BRC review of the draft framework were as follows:

1. Read the draft framework.

2. Review the performance expectations and performance indicators for each competency according to the bias review guidelines for test frameworks.

3. Discuss the framework's performance expectations and performance indicators for each competency with the other reviewers. Determine, as a group, whether the framework needs any revision according to the bias review guidelines.

4. If the framework needs revision, work as a group to suggest revisions according to the bias review guidelines.

5. The facilitator confirms the committee's determinations regarding the framework and records any committee revisions in a master copy of the Framework Review Booklet.

6. Each BRC member then signs the front cover of the master copy of the Framework Review Booklet to confirm that it is a true, complete record of the decisions of the committee.

Outcomes. As the framework was reviewed by BRC members, the facilitator documented any BRC recommendations for revisions, collected signatures from all BRC members to indicate that the documentation is complete, and presented the comments to the Content Advisory Committee (CAC). All four members of the BRC workgroup charged with reviewing the ALST Framework were satisfied that the framework was not biased and therefore required no bias-related revisions. The BRC members confirmed their decisions by signature. Following the BRC and CAC meetings, the results of both sets of reviews were submitted to NYSED, which made final decisions concerning test content.

Content Advisory Committee (CAC)

Function. The main function of the ALST CAC was to review test materials for appropriateness (including content accuracy, significance, job-relatedness, and freedom from bias), to incorporate bias-related comments (if any) from the Bias Review Committee, and to make final recommendations to the NYSED regarding test materials. Prior to starting its review, the CAC went through an orientation during which the members were trained on the review procedures. The ALST CAC was convened multiple times during the framework review and item review stages of the test development process.

Recruitment and Selection. The CAC for the ALST includes practicing New York State public school educators who hold permanent or professional certification in New York State and New York State educator preparation faculty (including education faculty and arts and sciences faculty). Since the ALST is taken by candidates seeking certification in multiple teaching areas, individuals certified and practicing in any of the fields for which the test is required were eligible to participate. In addition, the CAC members were recruited to include representation of the regional, gender, and ethnic diversity of New York State.

Several targeted recruitment efforts were conducted by the NYSED and Pearson in order to recruit and select qualified public school educators and educator preparation faculty to serve on the ALST CAC. The following two primary means were used to recruit potential members for the CAC: (1) large-scale distribution of nomination and application information to education professionals throughout New York State, and (2) targeted contact with educators identified through various means, including NYSED curriculum specialists and professional organizations. In addition, committee recruitment information was shared at professional conferences and meetings in New York State. NYSED reviewed the applications and selected individuals to serve on the CAC based on a review of their qualifications (e.g., educational position, professional activities, and representation of the diverse nature of New York State).

Composition. The ALST CAC is composed of 27 members including both certified and practicing New York State educators and educator preparation faculty. The ALST CAC included practicing and permanently certified public school educators and college and university faculty, including those teaching undergraduate or graduate education or arts and sciences courses in which prospective educators were enrolled.

Each time a meeting of the ALST CAC is convened, all members of the Committee are invited to participate in the review of ALST test materials. Refer to Appendix K for a description of the characteristics of the CAC members.

Content Review of the ALST Framework

The ALST Content Advisory Committee (CAC) systematically reviewed the draft ALST Framework to ensure that the test materials appropriately define the content of the test and meet the review guidelines specified for the ALST. The role of the CAC is to review test materials for appropriateness, including content accuracy, significance, job-relatedness, and freedom from bias.

NYSED and Pearson convened a meeting of the ALST CAC on April 19, 2012, at the Pearson office in Malta, New York. A total of 14 ALST CAC members participated in this meeting (including both certified and practicing New York State educators and teacher educators, as previously described in this chapter) to review the ALST Framework for accuracy and appropriateness.

Introduction, Orientation and Training. Staff from NYSED and Pearson conducted an orientation to the review session. During the orientation, the CAC members received information on the background and purpose of the ALST and the current CAC meeting; the review guidelines; and the procedures and materials that would be used in the framework review process.

The Pearson facilitator worked with the ALST CAC throughout the review process to clarify procedures and materials, address testing issues, facilitate discussions throughout the review, and keep the master copy of the CAC recommendations. All efforts were made to ensure that everyone had a voice in the review and that consensus regarding the test materials was being reached.

Materials. Each CAC member registered upon arrival by completing a Sign-in Sheet and a Confidentiality Agreement. CAC members received the following materials:

1. Orientation Manual
2. Framework Review Booklet
3. Framework Review Confirmation Form
4. Content Correlation Table

Review guidelines. CAC members were instructed to ask themselves a set of organized questions when reviewing the content of the ALST Framework. The questions related to: Program Purpose, Organization, and Inclusiveness.

> PROGRAM PURPOSE. Is the framework consistent with the purpose of the NYSTCE (i.e., to determine whether prospective educators have the knowledge and skills to perform the job of an educator in New York State)?
>
> ORGANIZATION. Is the framework organized in a reasonable way?
>
> INCLUSIVENESS. Is the content of the framework complete? Does the framework reflect the knowledge and skills an educator should have in order to perform the job of an educator? Is there any content that should be added?

The following questions were also considered by the CAC when reviewing the framework competencies, related performance expectations, and related sets of performance indicators within the framework: Significance, Accuracy, Freedom from Bias, and Job-Relatedness.

> SIGNIFICANCE. Do the competencies describe knowledge and skills that are important for educators to have?
>
> ACCURACY. Do the competencies accurately reflect the content, as it is understood by educators in the field? Are the competencies stated clearly and accurately, using appropriate terminology?
>
> FREEDOM FROM BIAS. Are the competencies free of elements that might disadvantage an individual because of her or his gender, race, ethnicity, nationality, national origin, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?
>
> JOB-RELATEDNESS. Do the competencies cover important knowledge and skills that an educator should have in order to perform the job of a New York State educator?

Review procedures. The CAC reviewed the draft ALST Framework to verify that the structure and the content of the framework met the review criteria. The procedures for the CAC review of the proposed assessment framework were as follows:

1. Read the framework and review the structure of the framework, as indicated by the Test Design, according to the framework review guidelines.

2. Review each competency and related performance expectations and set of performance indicators according to the framework review guidelines.

3. Discuss and determine, as a group, whether any framework components need revision according to the framework review guidelines.

4. Work as a group to make any revisions to the framework according to the framework revision guidelines.

5. The facilitator confirms the committee's determinations regarding the framework and records committee revisions in a master copy of the Framework Review Booklet.

6. Each individual committee member independently completes a Framework Review Confirmation Form.

Outcomes. The 14 members of the ALST CAC who participated in the Framework Review Conference reached consensus on revisions to the ALST Framework. The Pearson facilitator recorded the committee revisions in a master copy of the ALST Framework Review Booklet. A CAC representative reviewed the master copy of the Framework Review Booklet and signed the cover of the booklet to confirm that the master copy was a true, complete record of the decisions of the committee. All 14 CAC members individually completed and signed the Framework Review Confirmation Form, indicating that the ALST Framework met all of the framework review guidelines described above. There was a total of seventeen editorial and stylistic revisions on the ALST Framework recommended by the ALST CAC members: three revisions were recommended for the high-level description of the New York State educator who has the academic literacy skills necessary to teach effectively in New York State public schools; six were recommended for the *Competency 0001 Reading* Performance Expectations and Indicators; and eight were recommended for the *Competency 0002 Writing to Sources* Performance Expectations and Indicators. Example revisions include referring to the "New York State educator" instead of "the teacher" and taking out the adjective "persuasive" from the description of the type of text the educator is expected to engage with. All CAC-recommended revisions to the Framework were later reviewed by NYSED and fully implemented in the final version of the ALST Framework.

Conduct Content Validation Surveys

NYSED and Pearson conducted Content Validation (CV) Surveys in order to collect additional evidence confirming that the content of the Academic Literacy Skills Test (ALST) Framework is appropriate and job-related. A full report containing Content Validation Survey Results for the ALST can be found in Appendix O.

Results from the CV Survey reveal the degree to which NYS public school educators and educator preparation faculty consider each framework competency important for performing the job of an educator in New York State public schools.

The ALST test framework (test competencies and sets of performance indicators) was included in the CV Survey conducted with two populations: New York State public school educators and educator preparation faculty members from New York State education preparation programs. The purpose of the survey was to obtain judgments from New York State educators about:

- the importance of each ALST competency for performing the job of an educator in New York State public schools;
- how well each set of ALST performance indicators represents important examples of the knowledge and skills addressed by the competency; and
- how well the set of ALST competencies, as a whole, represents important aspects of the knowledge and skills needed to teach in New York State public schools.

Survey of New York State Public School Educators.

Sampling. In order to conduct the ALST CV Survey of public school educators for the ALST Framework, NYSED and Pearson selected a sample of educators who had initial or professional New York State certification and were practicing in an appropriate field to participate in the survey. To be eligible for inclusion in the sample, an educator had to be practicing or have practiced within the past year and be certified in the appropriate area for the 2010-2011 year, as indicated by the New York State Basic Educational Data System (BEDS) data files. A sample was selected of approximately 500 certified educators from the population of educators with appropriate assignment codes as identified by NYSED.

NYSED provided Pearson with the Personnel Master File (PMF) from the Basic Educational Data System (BEDS). In addition, Pearson downloaded the NYSED directory file from the NYSED website for school and district information. Pearson accessed the following information from the personnel file to select the public school sample:

- unique educator identifying number
- educator name
- teaching area
- teaching certification
- district and school codes

Pearson accessed the following information from the NYSED directory file:

- district codes
- district names and addresses
- school codes and names

Educators were randomly selected for the survey sample.

An advance notification letter was sent on NYSED letterhead to the schools of the randomly selected educators, in addition to an advance notification email from NYSED. The letter and email went to the school Principal, introducing Pearson, asking for cooperation with the survey, and providing notice that instructions would follow.

Survey Instrument. The survey instrument for the public school educators included the following elements, presented in a web-based survey:

1. General Directions: A brief overview of the New York State Teacher Certification Examinations and the purpose of the assessments was provided. In addition, general instructions for completing the survey were given.

2. Eligibility and Background Questions: Because survey recipients were expected to rate each test competency, the set of performance indicators corresponding to each test competency, and the entire set of competencies, it was necessary for recipients to be eligible to respond. To determine eligibility to complete the survey, recipients were asked to respond to two questions. Public school educators reviewing the content of the ALST Framework responded to the following two eligibility questions:

"Are you now a practicing educator or have you been a practicing educator during this or the previous school year in New York State public schools?"

"Do you currently hold a New York State teaching certificate or equivalent New York City license?"

To be eligible to participate in the ALST Content Validation Survey, an educator were required to respond "yes" to both questions. Recipients who answered "no" to either eligibility question were not able to proceed with the survey. Only ratings received from recipients who answered "yes" to both eligibility questions were included in the analysis of the survey data.

Respondents were also asked background questions to gather information about level of education, gender, ethnic/racial background, years of professional experience as an educator, grade level currently teaching, and certification level.

3. Survey Questions: Eligible survey participants were asked to rate the importance of each test competency, as well as the degree to which the set of performance indicators for each test competency represented important examples of the knowledge and skills addressed by the competency. Finally, they were asked to rate how well the set of competencies, as a whole, covered important aspects of the knowledge and skills required for performing the job of an educator in New York State public schools.

4. Collection of Job Importance Ratings: Survey recipients were provided with guidelines regarding their competency and performance indicator ratings. For the competency and

performance indicator ratings, they were asked to consider all aspects of the job of an educator in New York State public schools.

Rating question for competencies: Both public school educators and educator preparation faculty were asked to respond to the following question using the rating scale below:

"How important is the knowledge or skill indicated by this competency for performing the job of an educator in New York State public schools?"

1 = no importance
2 = little importance
3 = moderate importance
4 = great importance
5 = very great importance

Rating question for performance indicators: Survey recipients were also asked to rate the performance indicators related to each competency by answering the following question and using the following rating scale:

"How well does the set of performance indicators above represent important aspects of the knowledge and skills addressed by the competency?"

1 = poorly
2 = somewhat
3 = adequately
4 = well
5 = very well

Composite Rating: Survey recipients were asked to provide a composite rating of the complete set of competencies by answering the following question and using the following rating scale:

"How well does the set of competencies, as a whole, represent important aspects of the literacy knowledge and skills needed for performing the job of an educator in New York State public schools?"

1 = poorly
2 = somewhat
3 = adequately
4 = well
5 = very well

5. Collection of Comments: If survey participants indicated a rating of "1" or "2" (for competencies or performance indicators), they were asked to provide a comment on the appropriate screen, in order to provide the reason(s) for the rating. They were also asked to include any recommendations for changes or additional content.

Data collection. ALST CV surveys were administered as web-based surveys starting in November 2012 through February 2013.

Public school survey distribution emails were sent to Principals. Each email contained a survey access code for the randomly selected teacher, which the Principal was instructed to forward to the selected teacher.

As a result of the advance notification, there were some Principals whose emails were returned as "undeliverable" or Pearson received information that there were different contacts at particular schools. The Content Validation survey was conducted by paper letter and mail for this subset of schools. The envelopes were addressed to Principals, and contained a cover letter and letters containing survey access codes to distribute to the selected educators.

Approximately one month after the initial distribution, Pearson sent follow-up emails to Principals containing survey access codes for educators who had not responded. A similar follow-up mailing was sent to the Principals who had received the distribution by paper letter, including letters with survey access codes for educators who had not responded.

Returned and eligible surveys. Out of 500 surveys distributed to New York State public school educators, 223 were returned that were eligible for use in the data analysis. This return rate yielded a data set of a size comparable to similar statewide educator survey efforts across the country and enabled a sample of educators with an array of characteristics (see Appendix O).

Survey of New York State Educator Preparation Faculty.

Sampling. In order to conduct the ALST CV Survey of educator preparation faculty for the ALST Framework, NYSED and Pearson selected a sample of New York State educator preparation faculty. Educator preparation faculty at colleges and universities in New York State that prepare teacher candidates were eligible to participate in the faculty surveys.

Pearson generated a list of 122 eligible institutions, including information such as number of affiliated NYSTCE examinees in the testing fields and/or number of program completers prepared in relevant fields. Faculty from a total of 50 institutions were randomly selected to be distributed the ALST faculty surveys.

Survey Instrument. The survey instrument for the educator preparation faculty included the following elements, presented in a web-based survey.

1. General Directions: A brief overview of the New York State Teacher Certification Examinations and the purpose of the assessments was provided, and general instructions for completing the survey were given.

2. Eligibility and Background Questions: Survey recipients were asked to rate each test competency, the set of performance indicators corresponding to each test competency, and the entire set of competencies. For the responses to be accurate and meaningful, it was necessary for recipients to be eligible to respond. Educator preparation faculty reviewing the content of the ALST Framework responded to the following eligibility question:

"Have you, during this or the previous year, taught courses that may be taken by students preparing to become educators?"

Recipients who replied "yes" to the eligibility question were asked to complete the survey. Only ratings received from recipients who answered "yes" to the eligibility question were included in the analysis of the survey data.

Respondents were also asked background questions to gather information about level of education, gender, ethnic/racial background, years of experience as a faculty member, and the college/department of primary appointment.

3. Survey Questions: Eligible survey participants were asked to rate the importance of each test competency, as well as the degree to which the set of performance indicators for each test competency represented important examples of the knowledge and skills addressed by the competency. Finally, they were asked to rate how well the set of competencies, as a whole, covered important aspects of the knowledge and skills required for performing the job of an educator in New York State public schools.

4. Collection of Job Importance Ratings: Survey recipients were provided with guidelines regarding their competency and performance indicator ratings. For the competency and performance indicator ratings, they were asked to consider all aspects of the job of an educator in New York State public schools.

> Rating question for competencies: Both public school educators and educator preparation faculty were asked to respond to the following question using the rating scale below:
>
> "How important is the knowledge or skill indicated by this competency for performing the job of an educator in New York State public schools?"
>
> 1 = no importance
> 2 = little importance
> 3 = moderate importance
> 4 = great importance

5 = very great importance

Rating question for performance indicators: Survey recipients were also asked to rate the performance indicators related to each competency by answering the following question using the following rating scale:

"How well does the set of performance indicators above represent important aspects of the knowledge and skills addressed by the competency?"

1 = poorly
2 = somewhat
3 = adequately
4 = well
5 = very well

Composite Rating: Survey recipients were asked to provide a composite rating of the complete set of competencies by answering the following question and using the following rating scale:

"How well does the set of competencies, as a whole, represent important aspects of the literacy knowledge and skills needed for performing the job of an educator in New York State public schools?"

1 = poorly
2 = somewhat
3 = adequately
4 = well
5 = very well

5. Collection of Comments: If survey participants indicated a rating of "1" or "2" (for competencies or performance indicators), they were asked to provide a comment on the appropriate screen, in order to provide the reason(s) for the rating. They were also asked to include any recommendations for changes or additional content.

Data collection. ALST CV surveys were administered as web-based surveys in November 2012 through February 2013.

Before the opening of the Content Validation Survey, an advance notification letter was sent on NYSED letterhead, in addition to an advance notification email from NYSED. The distributions were sent to contacts at institutions with approved teacher preparation programs introducing Pearson, asking cooperation with the survey, and providing notice that instructions would follow.

Educator preparation faculty distribution emails were sent from Pearson to contacts at institutions. Institution contacts were either provided by NYSED or were Pearson score reporting contacts for the NYSTCE. The emails contained instructions on how to select eligible faculty members for each particular field. Institution Contacts were asked to forward survey access codes to faculty. Each institution received one survey access code per field.

Approximately one month after the initial distribution, Pearson sent follow-up emails containing survey access codes to contacts at institutions, encouraging them to follow up with eligible faculty who had not yet responded to the survey. In January 2013, in an effort to collect additional data from the faculty population, Pearson sent survey distribution emails to contacts at institutions for several fields, including the Academic Literacy Skills Test.

Returned and eligible surveys. Out of 112 surveys distributed to the NYS educator preparation faculty, 63 were returned and were eligible for the use in the data analysis. This return rate yielded a data set of a size comparable to similar statewide surveys of educator preparation faculty across the country and enabled a sample of faculty with an array of characteristics (see Appendix O).

Analysis of Content Validation Survey Results. Pearson analyzed the eligible responses to the surveys provided by both groups of respondents (i.e., public school educators and educator preparation faculty). Blanks were treated as missing data. Detailed results can be found in the final CV Survey Results Report in Appendix O; the section below provides a brief description and summary of the results for each report.

Background information. The demographic summary report summarizes the responses to the eligibility and background information questions in each survey. These summaries include the following:

- absolute frequencies—number of individuals selecting each response option, including nonresponses (blanks)
- relative percent—the percent of individuals selecting each response option, including nonresponses, rounded to one decimal place
- adjusted percent—the percent of individuals selecting each response option, excluding nonresponses, rounded to one decimal place

Data reports. Pearson prepared the following summary reports for ALST:

1. Rating Summary Report: The report provides statistics for each of the rating questions:

- Rate the competency
- Rate the set of performance indicators
- Overall composite

The statistics were calculated across all competencies based on the ratings for each test framework. The statistics include:

- Number of respondents
- Arithmetic mean of competency ratings, rounded to two decimal places
- Standard deviation of ratings, rounded to two decimal places
- Standard error of the mean, rounded to two decimal places

2. Ratings by Competency and Set of Performance Indicators Report: The report summarizes the ratings given to each competency and a set of performance indicators. The report provides the following information for each competency and a set of performance indicators:

- Number of respondents
- Arithmetic mean of the ratings, rounded to two decimal places
- Standard deviation of ratings, rounded to two decimal places
- Standard error of the mean, rounded to two decimal places
- Distribution of responses in percent for each response option (1-5) and no response

3. Comparison Reports: Where sample sizes permit, Pearson produced comparison group reports for the public school educator samples. The report lists mean importance ratings of each competency by group (e.g., by race/ethnicity). For classifications such as race/ethnicity, respondents were classified into groups according to self-reported information from the survey.

This report flags any competency that produces statistically significant results at the 0.05 level of significance when comparing mean competency ratings of any group of educators (e.g., Female) within the total group to mean competency ratings for the total group minus the focus group and where the mean importance rating for the focus group is less than 3.0. Focus groups to be evaluated in the comparison report include educators who self-report their gender as Female and those educators who self-report their ethnic or racial background as Black (not of Hispanic origin), Hispanic, or Other. Comparisons were not produced for any group with a size of fewer than 25 respondents, which is an industry-accepted threshold for a minimum sample size required for conducting statistical comparison analyses.

Comments. If survey participants indicated a rating of "1" or "2" (for competencies or performance indicators), they were asked to provide a comment, in order to provide the reason(s) for the rating. They were also asked to include any recommendations for changes or additional content. NYSED staff and Pearson testing specialists reviewed all collected comments and determined that no revisions to the ALST Framework were needed.

Summary of Results. All mean composite ratings for the overall sets of competencies were above 3.0, of moderate to very great importance for performing the job of an educator, for the public school sample and educator preparation faculty. All test competencies and performance indicator

sets in each test framework received mean ratings of more than 3.0, representing the competency adequately to very well, on the five-point scale from the public school educator sample and educator preparation faculty. No significant group differences were identified in the competency importance ratings, performance indicator set ratings, or the composite ratings in the public school educator survey sample.

The results of the Content Validation Surveys confirmed the appropriateness of the sets of competencies and associated performance indicators contained in the ALST frameworks. The results indicated that both public school educators and educator preparation faculty considered the competencies to be important for performing the job of an educator in New York State public schools.

The Content Validation survey results confirm that the set of competencies are important job tasks that all teachers must perform and that the competencies are appropriate to be assessed. Furthermore, the results provide support that the set of performance indicators included for each competency provides important examples of the knowledge and skills addressed by the competency.

New York State Educator Job Analysis

Establishing a strong connection between the job-related tasks performed by New York State educators and the content of the Academic Literacy Skills Test has been integral to the ALST test development process, starting with the initial surveys conducted during the development of the NYS Teaching Standards (see Chapter 1 for the detailed description). The importance of the literacy skills, as measured by the ALST, to performing the job of an educator is further demonstrated by the results of the New York State Educator Job Analysis (JA) study which is described in detail in this section of the report.

The goal of the New York State Educator Job Analysis (JA) study was to collect empirical evidence of the logical chain from critical tasks performed by educators in New York State, to the knowledge, skills, abilities, and other characteristics (KSAOs) (referred to as "competencies" in the ALST Framework and in this report) required for performing those tasks, to the inclusion of critical competencies in the ALST Framework. To this end, Pearson contracted the Human Resources Research Organization (HumRRO) to design and execute the study. Working in close collaboration with Pearson and NYSED at every step of the process, the HumRRO team consisted of industrial-organizational psychologists experienced in conducting high-stakes job analysis and assessment development studies, including development of certification examinations (see description of the project team on p. 2 in Appendix M). The summary of the JA study for the ALST is provided below. Please refer to the New York State Educator Job Analysis report for detailed information on the job analysis study (Appendix M).

The team conducting the JA study ensured that all activities were consistent with job analysis best practices and professional guidelines regarding development of employment and

certification examinations. All aspects of this study were conducted in accordance with the Standards for Educational and Psychological Testing (1999), as well as the Principles for the Validation and Use of Personnel Selection Procedures: 4th edition (SIOP, 2003), and Uniform Guidelines on Employee Selection Procedures (1978). New York State educators nominated for their expertise served as subject matter experts (SMEs) throughout the process (see Appendix M for the detailed description of the SMEs characteristics). They played a key role in each activity, ensuring that the materials were accurate and consistent with the minimum knowledge, skills, and abilities a new teacher needs to improve student learning in New York State.

The New York JA Study for the ALST consisted of three integrated steps: (1) delineation of critical tasks performed by the New York State educators, (2) compilation of evidence on the importance of various KSAOs for the job of an educator, (3) "linkage" between the KSAOs identified as important in step (2) and the KSAOs (or competencies) measured by the ALST. The results from each one of these activities are summarized below.

The first step was delineation of critical tasks performed by the New York State (NYS) educators and is the first link in the logical chain from occupational requirements to exam content. Using state and national standards, educator tasks listed in the U.S. Department of Labor's O*NET database, as well as their expertise as the basis, the NYS SMEs generated a draft list of tasks important for performing the job of an educator in the NYS (See Appendix M, Volume II). The task lists were inserted in standardized task surveys and administered to representative samples of currently certified and practicing educators in New York State, who rated the importance of each task on a 5-point Likert scale (ranging from 1 = minor importance for effective job performance to 5 = extremely important for effective job performance). These data were used to evaluate the relative importance of various tasks and to identify tasks that can be considered "critical" for the occupation of an educator. In order to identify "critical" tasks, HumRRO created a criticality threshold based on the importance rating and the time spent on rating each task, with more weight given to task importance (see p. 18 of Appendix M for the formula). Criticality value ranged from 3.0 to 15.0, and the criticality threshold was defined as follows:

- 90% or more of the respondents with non-missing data agree that they perform the task
- The task received a mean criticality rating of 8.0 or higher

Tasks falling above the threshold were considered critical. Table 3 lists the 34 tasks that exceeded the criticality threshold and were thus considered critical. Ultimately, the task information provides a basis for determining which competencies should be measured in the NYSTCE certification exams, including ALST.

Table 3. Teacher Tasks Identified as Critical

1. Aligns instructional plans with New York State learning standards, including Common Core
2. Establishes long-range goals and specifies the learning objectives and strategies to achieve them
3. Designs learning experiences that foster student understanding of key themes of the discipline
4. Creates developmentally appropriate unit and lesson plans and learning experiences that address students' learning differences and needs
5. Evaluates, selects, prepares, and modifies resources and materials for instruction to ensure comprehensiveness, content accuracy, and appropriateness for each learner
6. Designs learning experiences that connect students' prior knowledge to new content
7. Designs for differentiated instruction that reflects the diverse life experiences, strengths, and learning needs of all students
8. Adjusts lesson plans and instruction using a variety of strategies to address the needs of each student
9. Demonstrates knowledge of the New York State learning standards, including Common Core, and their application throughout instruction
10. Differentiates instruction and includes opportunities for students to achieve learning goals and objectives in a variety of ways
11. Provides students with essential questions and/or clear, measureable learning objectives
12. Provides and ensures understanding of clear oral and/or written directions, procedures, and expectations
13. Communicates and upholds high expectations for all students
14. Uses a variety of instructional and communication strategies, including technology, during instruction
15. Adjusts the pace and focus of instruction, as well as method of delivery, based on students' progress, comments, and questions
16. Equips students with strategies to persevere with challenging tasks
17. Uses creative and innovative approaches to learning
18. Observes and evaluates students' performance, behavior, social development, and general physical well-being
19. Assesses students to determine prior knowledge, skills, and understandings and to identify misconceptions
20. Uses technological tools to improve teacher productivity, keep accurate records, and enhance communication
21. Establishes, communicates, and upholds clear standards and high expectations for student learning and behavior
22. Models caring, respectful, and inclusive interactions toward all students
23. Develops, implements, and adapts routines and procedures to manage activities and transitions in a time-effective manner to achieve learning goals

Table 3. Teacher Tasks Identified as Critical (cont.)

24. Creates an environment that celebrates student accomplishments
25. Takes decisive and appropriate action to address problematic student behavior in the interest of safety
26. Maintains confidentiality regarding student records and information
27. Reports instances of child abuse, safety violations, bullying, and other concerns in accordance with laws, regulations, and policies
28. Shares information and best practices with colleagues
29. Demonstrates honesty, integrity, ethical conduct, and confidentiality when interacting with students, families, colleagues, and the public
30. Advocates to address the needs of all students
31. Reflects on feedback to inform goals for own professional growth
32. Reads professional literature and engages in ongoing professional growth and development
33. Builds and updates own academic skills and knowledge to adapt to higher academic standards and expectations for student achievement
34. Models or emulates practices exhibited by highly effective teachers

Note. Tasks are listed in the order in which they appeared in the task survey.

The second step was compiling existing evidence on the importance of various competencies for the educator occupations, supplemented by evaluations of the importance of (a) detailed "knowledge facets" (referred to as "performance indicators" in the ALST Framework and in this report) underlying the ALST, and (b) broader skills, abilities, and work styles. This is the second link in the logical chain from occupational requirements to exam content. All competencies (and the performance indicators) measured by the ALST were found to be critical for performing the job of an educator. Table 4 lists the competencies measured by the ALST, followed by mean importance ratings from random samples of New York State (a) educators and (b) educator preparation faculty (refer to Appendix M for characteristics of the committee members). It also shows the number of performance indicators associated with each competency and the number receiving a mean importance rating higher than 3.5 (on a 5-point scale) from focus group SMEs.

Table 4. Importance Ratings of Competencies and Performance Indicators for the ALST

Competency	Mean Importance Rating		# of Performance Indicators (PIs)	# of PIs with Mn Imp > 3.5
	Educators	Ed Prep Faculty		
Reading	4.60	4.85	12	12
Writing to Sources	4.28	4.62	15	15

Note. N = 194-224 certified educators and 33-60 educator preparation faculty evaluated competency importance; 24-25 teacher SMEs provided performance indicator judgments. Rating scale anchors are 1 = no importance, 2 = little importance, 3 = moderate importance, 4 = great importance, 5 = very great importance. Mn Imp = mean importance.

In summary, certified educators and educator preparation faculty provided an average rating above 4 (4 = great importance) to both ALST competencies. Furthermore, all 12 performance indicators subsumed under the *Reading* competency and all 15 performance indicators subsumed under the *Writing to Sources* competency received an average rating of above 3.5 (3 = moderate importance). Therefore, all the competencies (and associated performance indicators) measured by the ALST are critical for performing the teacher occupation.

The third and final step was a "linkage" exercise in which SMEs (see Appendix D for characteristics of the SMEs) evaluated the importance of the competencies identified in the preceding step for each of the tasks included in the task lists. These judgments are used to ensure that only competencies required for performing critical job tasks are included in the ALST. Thus, they provide the third link in the logical chain from occupational requirements to the ALST content.

The rating scale used by SMEs essentially incorporated two judgments. First, SMEs had to determine if a particular competency is needed at all to perform a particular task. If not, then they assigned a rating of "1" and moved to the next cell in the matrix. In cases where SMEs determined that the competency is needed to perform a particular task, they next had to determine whether the competency is important or extremely important for performing the task, which dictated the choice between a rating of "2" and a rating of "3." An example portion of the teacher linkage matrix (with the rating scale in the top left column) is shown in Figure 2.

1 = Not Important. The KSAO is **not needed** to perform the task. **2** = Important. The KSAO is **needed** to perform the task. **3** = Extremely Important. The KSAO is **essential** to perform the task.	Competency (KSAO)	
	1	2
A. Plans instruction		
1. Aligns instructional plans with New York State learning standards, including Common Core		
2. Aligns instructional plans with professional national learning standards (e.g., IRA, NCTM)		
3. Develops instruction that reflects an understanding of the school community		
4. Establishes long-range goals and specifies the learning objectives and strategies to achieve them		
5. Designs learning experiences that foster student understanding of key themes of the discipline		

Figure 2. Example portion of the teacher linkage matrix

Linkage judgments collected for the competencies measured by the ALST indicated that each competency is linked to more than one critical job task, and most are linked to several critical job tasks. A "link" was defined as any competency-task combination for which at least 75% of the SMEs assigned a rating of 2 or 3. Table 5 shows the number of teacher tasks linked to each ALST competency across all 105 tasks and the 34 critical tasks. The linkage judgments provide yet another way to ensure that the ALST frameworks measure appropriate (i.e., critical) competencies.

Table 5. Number of Teacher Tasks Linked to Each Competency of the ALST

Competency	# Tasks for which ≥ 75% of SMEs Assigned a Rating of 2 or 3	
	Across All 105 Tasks	Out of 34 Critical Tasks
Reading	66	20
Writing to Sources	61	20

Note. $N = 9$ Teacher SMEs. Rating scale anchors are: 1 = not important. The KSAO is not needed to perform the task. 2 = important. The KSAO is needed to perform the task. 3 = extremely important. The KSAO is essential to perform the task.

In summary, every ALST competency is linked to multiple tasks identified as important for performing the job of an educator in the New York State. Furthermore, every ALST competency is linked to 20 tasks identified as *critical* job-related tasks. These results support the content validity of the ALST by showing that every competency measured by the ALST is needed to perform critical job tasks.

As part of the development of the ALST Framework, a thorough Job Analysis study was conducted in accordance with professional guidelines and best practices to further assure the job-relatedness of the ALST . The JA information provided foundational support for identifying the characteristics required to perform critical teacher tasks and built a logical chain from critical tasks performed by the NYS educators to critical knowledge, skills, abilities, and other characteristics (KSAOs) (or competencies) required to perform those tasks, to measurement of these competencies on the ALST. Documentation of the full logical chain is the core of content validation, a professionally accepted practice for demonstrating that the content of an examination accurately and representatively reflects only KSAOs critical for performing the target occupation.

Development of the ALST Items

Overview

After sufficient content-based validity evidence for the appropriateness and job-relatedness of the ALST Framework was collected, the process of developing items matched to the framework commenced. Analogous to the framework development and validation, several steps were taken to build content-based validity evidence for the items. First, Assessment Specifications (Appendix P) were created in accordance with the job-related competencies outlined in the ALST Framework and served as a foundation for creating prototype ALST items by NYSED and Pearson. Second, after a sufficient number of ALST items were created, they were taken to the Item Review Conference (see the detailed description later in this chapter). During this effort, a Bias Review Committee (BRC) reviewed the content of the ALST items and was charged with bias prevention. In addition, the BRC also reviewed ALST items for appropriateness and job-relatedness. Following the Bias Review, a Content Advisory Committee (CAC) reviewed the items for content appropriateness and job-relatedness, incorporated bias-related comments from the BRC, and made final recommendations to NYSED regarding revisions to the ALST items. After the items have been vetted by the committees, they were field-tested in order to collect evidence of their performance in the field. Next, a Marker Establishment meeting was conducted with New York State educators in order to identify a set of responses exemplifying each of the score points on the scoring rubric. First operational ALST test forms (and, subsequently, equivalent forms) were then assembled. Finally, a Standard Setting conference was conducted with the NYS educators in order to recommend performance standards for the ALST.

Assessment Specifications

The ALST Assessment Specifications provided guidelines for the development of the ALST item bank and included the Test Blueprint, Item Development Specifications, and Passage Specifications (see Appendix P), which were used to create item prototypes. The Assessment Specifications were developed by Pearson testing specialists (a team of full-time senior test development staff that included literacy experts, former educators, psychometricians, and leaders with decades of experience developing assessment materials for educator credentialing in several states across the country) in close collaboration with NYSED.

The Test Blueprint specifies a number and type of items matched to each of the ALST competencies (see Chapter 2 for a detailed description of the ALST Test Blueprint). It served as a guide for item writing as well as for test form assembly.

Item Development Specifications define criteria for item development that operationalize the content to be assessed, as defined by the ALST Framework. The preceding sections of this report have described the purpose and use of the ALST and how the framework structures the content in a manner consistent with the test's purpose and use. The ALST item development specifications further extend this series of links, by relating the types and characteristics of the

items to be developed to the content that would be assessed. The specifications operationalize how reading comprehension and analysis skills and written analysis and expression skills should be measured, in a manner that would reflect the New York State vision of what it means to be a literate person in the twenty-first century. The ALST is designed to measure whether candidates are capable of attentive, critical reading that reflects "wide, deep, and thoughtful engagement" with high-quality informational texts and "cogent reasoning and use of evidence that is essential to both private deliberation and responsible citizenship in a democratic republic."[15] These measurement goals are achieved via selected-response and constructed-response items carefully designed to assess two competencies: (1) *Reading* and (2) *Writing to Sources.* ALST Item Development Specifications were used to help ensure that the items developed for each competency in the framework have common key characteristics and employ a consistent approach toward the measurement of that content.

Passage Specifications were also guided by the reading and writing skills described in the New York State Teaching Standards and operationalized in the NY P-12 CCLS for ELA and Literacy. The ALST passages were carefully selected to include a wide range of literary and informational texts, from the 1700s to the present.

NYSED-approved Assessment Specifications were used to guide development of item prototypes for each type of ALST item (i.e., selected-response, focused- and extended-constructed response items). The prototype items were reviewed by Pearson content experts and NYSED. After NYSED approved item prototypes, development began for items that would later comprise the ALST item bank. The Assessment Specifications guided passage selection and item development by professional item writers experienced in the development of items for educator credentialing examinations. Senior test development specialists at Pearson with decades of experience developing assessment materials for educator credentialing in several states across the country (working closely with a team that included literacy experts, former educators, psychometricians, and others) served as editors for the ALST item banks.

Conduct Item Review Conferences

Bias Reviews of Test Items.

For the description of the function, recruitment, selection, and composition of the ALST BRC refer to the *Bias Review Committee (BRC)* section under *Conduct Framework Review Conferences* section of this chapter. For the description of characteristics of the BRC members, refer to Appendix K.

The ALST Bias Review Committee (BRC) systematically reviewed the draft ALST items. This review was conducted to ensure that the ALST draft items were free from bias. The comments

[15]Introduction, Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects, p. 3.

and recommendations of the BRC were then communicated to the Content Advisory Committee (CAC) and were then incorporated into the review of, and revisions to, the draft ALST items.

The item development effort required to assemble an item bank for ALST was so large that the items were developed and reviewed in phases. The Phase 1 NYSTCE BRC meeting at which ALST items were reviewed was conducted at the Pearson office in Malta, New York, on January 17, 2013. A total of 10 members of the BRC (including New York State educators and educator preparation faculty, as described previously in this chapter), reviewed the ALST test materials in assigned workgroups. Two workgroups were charged with reviewing the ALST items and associated stimuli, each consisting of five BRC members generally representative of the diversity and expertise of the full BRC (including both New York State educators and educator preparation faculty, as described previously in this chapter). One group reviewed a booklet of 176 selected-response items (22 passages with eight SRIs each), and another group reviewed a booklet of 81 constructed-response items (27 stimuli sets with three CRIs each).

Phase 2 of the NYSTCE BRC meeting at which the next set of the ALST items were reviewed was conducted at the Pearson office in Malta, New York, on November 4 and November 5, 2013. On November 4, 2013, a group of 12 BRC members reviewed a booklet of 144 selected-response items (18 passages with eight SRIs each). On November 5, 2013, a group of 11 BRC members reviewed a booklet of 36 constructed-response items (12 stimuli sets with three CRIs each). The next section provides a description of the meetings and the process used to review the ALST items.

Introduction, Orientation and Training. Staff from NYSED and Pearson conducted an orientation for the bias review session. During the orientation, BRC members received information on the background and purposes of the NYSTCE program, the purpose of the current meeting, and the review process for the test materials. Prior to starting their review, committee members were trained on best practices in detecting and preventing bias in tests and received a copy of *Fairness and Diversity in Tests* (2009). The training was conducted by Pearson testing specialists with many years of experience providing similar training and facilitation for several other bias review committees engaged in the review of assessment materials for educator credentialing programs across the country.

Each BRC workgroup was assigned a Pearson facilitator, who oriented committee members to their tasks, explained procedures and materials, addressed testing issues and ensured that appropriate progress was being made, and kept the master copy of the workgroup's recommendations. All efforts were made to ensure that everyone had an opportunity to share his or her opinion in the review and that consensus regarding the test materials was being reached.

Materials. Each BRC member registered upon arrival by signing a Sign-in Sheet and a Confidentiality Agreement. BRC members received the following materials:

1. Orientation Manual

2. Test Design and Framework
3. Item Review Booklets
4. Answer Key
5. *Fairness and Diversity in Tests*

Bias Review Guidelines for Test Items. BRC members were asked to focus on the following questions about bias-related issues to review the items:

Content Does the item contain content that disadvantages a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Language Does the item contain language that disadvantages a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Offense Is the item presented in such a way as to offend a person because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Stereotypes Does the item contain language or content that reflects a stereotypical view of a group based on gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background?

Fairness Taken as a whole, are the items fair to all individuals regardless of race, gender, cultural background, or other personal characteristics?

Diversity Taken as a whole, do the items include content that reflects the diversity of the New York State population?

Review Procedures. BRC members reviewed the items to verify that they met the bias review criteria. The procedures for the BRC review of the items were as follows:

1. Read the test materials (items and associated stimuli).

2. Review the test materials according to the bias review guidelines.

3. Discuss the test materials with the other reviewers. Determine, as a group, whether the test materials contain potential bias according to the bias review criteria.

4. If potential bias is found, work with the group to recommend a revision, using the bias review guidelines.

The facilitator confirmed the committee's determinations regarding the test materials and recorded committee recommendations on a master copy of the Bias Review Committee

Comment Form. At the completion of the review of items for a field, the Bias Review Committee members reviewed all the recommended changes to the items and verified by signing the master copy that the document was a fair and accurate record of the committee's decisions.

Summary of Outcomes. As a result of Phase 1 BRC (January 17, 2013), two bias-related revisions were suggested by each BRC workgroup for group of 176 selected-response items, and two bias-related revisions were suggested for the group of 81 constructed-response items. All participating BRC members signed off on these revisions (five in each workgroup). The suggested revisions were then relayed to the Content Advisory Committees, who incorporated revisions to address the BRC's bias-related comments.

As a result of Phase 2 BRC (November 4–November 5, 2013), nine bias-related revisions were suggested by the BRC workgroup for the group of 144 selected-response items, and one bias-related revision was suggested for the group of 36 constructed-response items. All participating BRC members signed off on these revisions (12 members on November 4, 2013, and 11 members on November 5, 2013). The suggested revisions were then relayed to the Content Advisory Committees (CAC), who incorporated revisions to address the BRC's bias-related comments.

Following the ALST BRC review, the facilitator documented the BRC recommendations for revisions in the Master Item Review booklet, collected signatures from all BRC members to indicate that the documentation was complete, and presented the comments to the Content Advisory Committee (CAC). The CAC was instructed to address all bias-related issues raised by the Bias Review Committee (BRC). Following BRC and CAC meetings, the results of both sets of reviews were submitted to NYSED, which made final decisions concerning test content.

Content Review of Test Items.

For the description of the function, recruitment, selection, and composition of the ALST CAC refer to the *Content Advisory Committee (CAC)* section under *Conduct Framework Review Conferences* section of this chapter. For the description of characteristics of the CAC members, refer to Appendix K.

The ALST Content Advisory Committee (CAC) systematically reviewed the ALST items and associated stimuli to ensure that the test materials appropriately reflect the important tasks of a teacher's job and the content of the framework and meet the review guidelines specified for NYSTCE tests. The role of the CAC is to review test materials for appropriateness, including match to framework and standards, accuracy, freedom from bias, and job-relatedness.

As with bias review, content review of ALST items was accomplished in two phases. Phase 1 of the NYSTCE CAC meeting at which the ALST items were reviewed was conducted at the Pearson office in Malta, New York, on January 22–24, 2013. A total of ten ALST CAC members, generally representative of the expertise and characteristics of the full CAC,

participated on all three days (including both certified and practicing New York State educators and educator preparation faculty, as previously described in this chapter) to review the ALST items for appropriateness, including match to framework and standards, accuracy, freedom from bias, and job-relatedness. Two item booklets were reviewed: a booklet with 176 selected-response items (22 passages with eight SRIs each), and another booklet with 81 constructed-response items (27 stimuli sets with three CRIs each).

Phase 2 of the NYSTCE CAC meeting at which the ALST items were reviewed was conducted at the Pearson office in Malta, New York, on November 5–7, 2013. A total of 15 ALST CAC members participated in this meeting on November 5 and November 7, 2013, and 16 ALST CAC members participated during the meeting on November 6, 2013 (including both certified and practicing New York State educators and educator preparation faculty, as previously described in this chapter), to review the ALST items for appropriateness, including match to framework and standards, accuracy, freedom from bias, and job-relatedness. Two item booklets were reviewed: a booklet of 144 selected-response items (18 passages with eight SRIs each) and a booklet of 36 constructed-response items (12 stimuli sets with three CRIs each).

Introduction, Orientation, and Training. Staff from NYSED and Pearson conducted an orientation to the review session. During the orientation, the CAC members received information on the background and purpose of the NYSTCE program, the purpose of the ALST and the current CAC meeting, and the review criteria, procedures and materials that would be used in the item review process.

The Pearson facilitator worked with the ALST CAC throughout the review process to clarify procedures and materials, address testing issues and ensure that appropriate progress was being made throughout the review, and keep the master copy of the CAC recommendations. All efforts were made to ensure that everyone had a voice in the review and that consensus regarding the test materials was being reached.

Materials. Each CAC member registered upon arrival by completing a Sign-in Sheet and a Confidentiality Agreement. CAC members received the following materials:

1. Orientation Manual including Item Review Criteria, ALST Prototype Items, ALST Match Guide, ALST Item Review Practice Session Materials, and Common Core Regents Items
2. Test Design and Framework
3. New York State P-12 Common Core Learning Standards for English Language Arts & Literacy (Excerpts)
4. Item Review Booklet
5. Answer Key
6. Item Rating Forms
7. Review Criteria.

Review guidelines. The following guidelines were used by the CAC when reviewing the ALST items and associated stimuli: *Match to Framework*, *Accuracy*, *Bias*, and *Job-Relatedness*.

I. Match to Framework

Each item and associated stimuli for the test should match the Test Framework. The items have been developed to match the Test Framework – in order to measure the knowledge, skills, and abilities that are described by the framework's competencies, performance expectations, and performance indicators.

Background knowledge of the subject referenced in the text should not advantage candidates. Questions should focus soundly on comprehension and analysis of the stimulus material.

Each item you review should measure the knowledge, skills, or abilities described by the competency described in the Framework. The item need not cover the entire range of content for the competency, but it should measure important content that is clearly linked to one or more of the performance indicators. For each item, there should also be a clear connection between the content of the item and the standards.

As you review the test materials, please confirm the following:

- The item measures knowledge, skills, or abilities described in the Framework.

II. Accuracy

Each item and associated stimuli must be accurate. As CAC members reviewed the items, they were asked to confirm the following:

- The content of the item is accurate.
  - The item is factually correct.
  - The item correctly represents the content presented in any stimulus material.
  - There is one unambiguously correct response.
- All parts of the item are clear.
  - The terminology is appropriate.
  - Language is in the clearest form to measure the knowledge, skills, or abilities assessed by the item.
  - The item is free of typographical and grammatical errors.
- The item is appropriately constructed.
  - The structure of the item is sound (e.g., response choices are consistent in length, structure, and language).
  - The wording of the item stem is free of clue-ins.
  - Distractors are plausible and do not introduce ambiguity.

III. Freedom from Bias

Each test item should be free of bias. There are two broad aims of bias prevention in test items:

1) To exclude from test items language, content, or stereotypes that might disadvantage or offend an examinee because of her or his gender, race, nationality, national origin, ethnicity, religion, age, sexual orientation, disability, or cultural, economic, or geographic background.
2) To include content that is fair and equitable for examinees, regardless of gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background – and that reflects the diversity of New York State.

The language and content of each item should be free of stereotypes, and should not potentially disadvantage or offend an individual because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background. Each item should be fair and equitable.

As you review each item, please confirm the following:

- The item is free of language, content, or stereotypes that might potentially disadvantage or offend an individual because of her or his gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background.
- The items, as a whole, are fair to all individuals regardless of gender, race, religion, age, sexual orientation, disability, or cultural, economic, or geographic background. As a whole, the items include content that reflects the diversity of the New York State population.

IV. Job-Relatedness

Each item and associated stimuli should measure important knowledge and skills that a certified New York State educator should have to perform her or his job in New York State schools. The content should reflect knowledge and skills that may be expected of an educator at the entry level for the associated educator certificate, rather than learned later on the job.

As you review each item, please confirm the following:

- The item measures knowledge and skills needed in order to perform the job of a New York State educator, as reflected in the Test Framework and standards.
- The item measures knowledge and skills at a level that is appropriate for the certificate(s) for which the test is a requirement, as reflected in the Test Framework and standards.

Consider content taught directly to students, as well as knowledge and skills used in planning, designing, and delivering instruction; managing the classroom environment; assessing the performance of students and the effectiveness of instruction; interacting with students, families,

other education professionals, and community members; and other responsibilities of a certified educator in New York State.

Item Review Procedures. The CAC reviewed the ALST items to verify that the ALST items and associated stimuli met the review criteria. The procedures for the CAC review of the ALST items and associated stimuli were as follows:

1. Independently read and respond to each item to be discussed (as would an examinee), then check your answers.

2. Read the competency within the Test Framework (including Performance Expectations and Performance Indicators) to which the item is linked.

3. Review the Item Match Guide and the element(s) of the Common Core Learning Standards to which the item is linked.

4. Review the item according to the content review guidelines for test items.

5. Discuss each item with the other reviewers. Determine, as a group, whether each item needs any revision according to the content review guidelines.

6. If an item needs a revision, work with the group to revise the item according to the item review criteria and the item revision guidelines.

7. The Pearson facilitator will confirm the committee's determinations regarding the item and record any committee revisions in a master copy of the Item Review Booklet.

8. Independently provide an item rating for the item on the Item Rating Form, according to the review guidelines. Use the Item Rating Procedures to record your ratings.

Item Rating Procedures. After each item was revised and approved by the CAC, each committee member provided an item-by-item judgment as to whether or not an item is appropriate to use on a test form. All judgments were recorded.

Materials. The following materials were used to record item-level judgments:

1. Item Review Booklet
2. Test Framework, Item Matching Guide, and Common Core Learning Standards
3. Content Review Guidelines for Test Items
4. Rating Forms

The following procedures were used for item ratings:

1. Respond to the question in the box below by marking "1" or "2" in Column 1 of the Item Rating Form.

   COLUMN I: Is the item valid?

   COLUMN I: VALIDITY

   YES = 1 NO = 2

   Rating an item Valid affirms that the item:

   - matches the Test Framework and Common Core Learning Standards
   - is accurate
   - is free of bias
   - is job-related

   If you responded YES, do not mark anything in columns II and III.

   If you responded NO, proceed to Step 2 to complete a rating in Column II.

2. If you mark "2" in Column I to indicate NO, the item is not valid, you must mark the reason in Column II of the Rating Form. Respond to the question in the box below by marking "M", "A", "B", or "J" in Column II of the Item Rating Form.

   COLUMN II: If the item is not valid, what is the reason?

   COLUMN II: REASON FOR RATING ITEM NOT VALID

   Record the reason(s) for rating the item "Not Valid" in Column II of the Rating Form by marking:

   M to indicate the item does not match the Test Framework and Common Core Learning Standards.

   A to indicate the item is not accurate

   B to indicate the item is not free of bias

   J to indicate the item is not job-related

   Leave Column III blank.

Outcomes. Following Phase 1 (January 2013) CAC review of the items, 3 CRIs were deleted and 9 CRIs were designated for subsequent review at the future IRC (out of total 81 CRIs reviewed); 16 SRIs were deleted (out of total 176 SRIs reviewed). Following Phase 2 (November 2013) CAC review of the items, 6 CRIs were deleted and 3 CRIs were designated for subsequent review at the future IRC (out of total 36 CRIs reviewed); 32 SRIs were deleted (out of total 144 SRIs reviewed).

Conduct Field Testing

Purposes of Field Testing. New York State educator certification candidates participated in the field testing of items for the Academic Literacy Skills Test (ALST). ALST items were field tested with this population in order to have a field test population that would represent the characteristics of the ALST operational testing population – which would also be taken by New York State educator certification candidates. Field-test performance provided information that was used to gauge expected operational test performance. Field-testing results were used to determine that the test items were appropriate, clear, reasonable, and had acceptable statistical and qualitative characteristics. Additionally, field-test results helped to inform the implementation of the operational administration of the ALST (e.g., testing time required).

Field-test strategies. Multiple strategies were employed for conducting field testing of ALST items in New York State. These strategies are described below.

1. Field testing selected-response items on stand-alone field-test forms. Field testing ALST items on stand-alone field-test forms enabled the collection of empirical evidence that test items were properly functioning and were demonstrating sound psychometric characteristics prior to the creation of the first operational-test forms. Multiple ALST field-test forms were created, in order to collect data on a sufficient number of items to enable the sound construction of initial operational-test forms. 20 different field test forms were administered across two rounds of stand-alone field testing, resulting in over 2500 total ALST field test forms taken by participants. At least 73 field test participants responded to each selected-response item administered.

Field-test participants were permitted to complete multiple field-test forms, but were not permitted to repeat the identical form. ALST field-test forms, like operational ALST test forms, were administered at computer-based testing (CBT) centers, using CBT software to ensure the delivery of unique field-test forms to candidates taking more than one form. Field-test forms were randomly assigned to participants. Field-test participants did not receive a Pass/Did Not Pass status on the exam and were not provided a total test score report.

As previously mentioned, field testing of ALST items occurred in two rounds: the first round occurred between May 13, 2013, and June 3, 2013, using field-test forms that included both selected-response and constructed-response items; and the second round occurred between August 26, 2013, and September 20, 2013, using field-test forms that included constructed-response items only. Conducting field testing in two rounds helped to ensure that a sufficient number of constructed-response items were field tested prior to their use on operational-test forms.

In order to help ensure that candidates' performance on field-test forms would be comparable to candidates' performance on operational-test forms, several "control" items from the operational NYSTCE exams with known performance characteristics were included on field-test forms. Item

characteristics were then compared from the field test and operational administrations to check for comparability.

2. Field testing selected-response items on operational test forms. Field testing of ALST selected-response items was also conducted by embedding items as "non-scorable" items on operational ALST forms. Operational ALST forms are composed of 80% (32 out of 40) selected-response items that have been previously field tested and are thus designated as "scorable" – that is, contributing to the total score. The remaining selected-response items (20%, 8 out of 40) are designated as non-scorable – that is, they do not contribute to the total score. These non-scorable items are embedded on the operational-test forms in order to collect data on the psychometric characteristics of the items prior to their use as scorable items. Selected-response items are being continuously field tested in this manner. Candidates are unaware which items are non-scorable; therefore, item performance can be expected to mirror operational performance. Selected-response items embedded as "non-scorable" items on operational test forms are being monitored continuously, using similar flagging criteria as specified above.

Recruitment of participants. Candidates enrolled in New York State educator preparation programs were encouraged to participate in the field testing of ALST items on stand-alone field-test forms. In order to recruit appropriate candidates to complete field-test forms, NYSED and Pearson employed the following recruitment strategies:

- A news announcement was posted on the NYSTCE website inviting eligible candidates to participate in field testing
- All candidates registered for any NYSTCE test were notified via email regarding field-testing opportunities
- Educator preparation program representatives were notified via email regarding field-testing opportunities and asked to distribute this information to potential field-test participants
- Content Advisory Committee (CAC) members were notified via email regarding field testing and asked to distribute this information to potential field-test participants

Participants registered for the ALST field-test sessions on the NYSTCE website and chose their preferred time and location for field testing.

Field test incentives. NYSED and Pearson offered the following incentives to each participant who completed a field-test form:

- $100 electronic coupon code for Penguin Books OR $50 voucher that may be applied to future NYSTCE test fees; and
- Viewing raw score results for selected-response items at the completion of the field-test session.

In addition to the incentives above, candidates whose scores on the selected-response items placed them in the top 10 percent of all candidates participating in the field test were entered into a drawing to win one of two available iPads and two $100 VISA or MasterCard gift cards.

Construction of field-test forms. For the first round of field testing (May 13–June 3, 2013), each ALST field-test form was constructed to contain the following:

- Two sets of reading prompts, each with eight selected-response items (16 SRIs total per field-test form)
- One set of constructed-response items: one extended-response assignment and two focused-response assignments (three CRIs total per field-test form)
- Three selected-response "control" items from an existing NYSTCE test with known performance characteristics from operational data

A total of 96 ALST selected-response items and six CRI sets (12 focused-response and six extended-response items) were field tested during the first round.

For the second round of field testing (August 26–September 20, 2013), each ALST field-test form was constructed to contain the following:

- One set of constructed-response items: one extended-response assignment and two focused-response assignments (three CRIs total per field-test form)

A total number of 14 CRI sets (28 focused-response and 14 extended-response items) were field tested during the second round.

Procedures for administration of field-test forms. Field testing was available as computer-based testing at the Pearson testing centers, including Pearson Professional Centers (PPCs) and Pearson Authorized Test Centers (ATCs) during both rounds of field testing. Field testing was available to candidates during the same days and times that other NYSTCE computer-administered tests were available.

In addition to completing a field-test form, candidates were also asked to complete a Computer-Based Testing (CBT) tutorial and a Non-disclosure Agreement (NDA) prior to testing and a questionnaire at the end of testing.

During the first round of ALST field testing, candidates had a total of 160 minutes to complete the field-test form. During the second round of ALST field testing candidates had a total of 120 minutes to complete the field-test form.

The field-test forms were administered under the same conditions as operational test sessions. Participants responded to test items under conditions of quiet, confidentiality, and test security. Participants were subject to the same identification requirements and procedures as candidates

taking operational tests and were not permitted access to unauthorized aides, such as reference materials or notes.

Scoring. Field-test responses to the selected-response items were scored electronically using the appropriate test form answer keys. Trained and calibrated NYSTCE scorers, composed of current and former New York State educators and educator preparation faculty, scored the responses to the constructed-response items using the rubric that was developed for the ALST items. Scorers' comments were considered when reviewing the field-test results for the constructed-response items. For an explanation of the focused holistic scoring approach used to score ALST CRIs and for a discussion of the development and use of the rubric and related scoring materials, please see the Scoring Section in Chapter 4.

Data analysis.

1. Selected-response items. The following field-test analyses were conducted for the ALST selected-response items:

- individual item p-values (percentage of participants answering the item correctly)
- item point-biserial correlation (item-total correlation or an index of differentiating power of an item)
- distribution of participant responses (percentage of participants selecting each response option, including "no response")
- mean score by response choice (average score on the selected-response set achieved by all participants selecting each response option)

2. Constructed-response items. The following field-test analyses were conducted for the constructed-response items:

- mean score
- standard error of measurement
- standard deviation
- distribution of scores
- analysis of variance (ANOVA) to detect item main effects differences (provided that the number of responses is greater than or equal to 25)
- analysis of variance (ANOVA) for item-by-participant group interactions (provided that the number of responses is greater than or equal to 25)
- analysis of the comparable difficulty of each item set on constructed-response item forms using Tukey's Studentized Range (Honest Significant Difference) Test for Total Score
- qualitative analyses of participants' and scorers' comments

Summary of Field Test Outcomes.

1. Selected-response items. Selected-response items with appropriate statistical characteristics, based on the field-test analyses, were included in the final item bank and were available for operational use on an ALST test form.

As previously mentioned, at least 73 field test participants responded to each selected-response item. This number of participants is comparable to numbers participating in similar field test efforts across the country for states' educator credentialing assessments. Of the items included on field test forms, the percentage of field test participants who correctly answered each item (p-value) ranged from 0.11 to 0.95. The item-to-test point-biserial correlation (item-total correlation or an index of differentiating power of an item) for the items ranged from 0.01 to 0.63.

Selected-response items were flagged for further content review and/or additional field testing if one or more of the following conditions applied:

- the p-value was less than 0.30
- the item-to-test point-biserial correlation was less than 0.10 and the p-value was less than 0.90 (provided the number of respondents was greater than or equal to 25)
- bimodal distribution of responses (i.e., one of the incorrect response options, as well as the correct response, attracted a large number of candidates)
- nonmodal distribution of responses (i.e., incorrect and correct responses options attracted a similar number of candidates)

Item statistics were used in conjunction with further item review by content experts to recommend whether a flagged item should be (a) deleted from the item bank, (b) reviewed and revised, as necessary, by ALST CAC members, (c) subsequently field tested as nonscorable items on the operational-test forms. Each recommendation was made on an item-by-item basis and was informed by careful review of the item statistics and item content.

As a result of the first round of field testing on field-test forms and subsequent data analysis and content review, out of 96 selected-response items field-tested, 85 items were identified for inclusion on the operational test form, four items were deleted, and seven were identified for further review and revision by the Content Advisory Committee and subsequent re-field-testing on operational test forms.

The ALST items are not subject to statistical bias detection techniques, such as tests for differential item functioning (DIF), due to the following unique circumstances and constraints surrounding test development. The relatively low sample sizes available for field test data do not allow for meaningful statistical analysis for DIF detection. Typically, the rule of thumb used as a standard industry practice is that for both the focal group and the reference group, the minimum sample size per group should be 200. If the sample size for either of the two groups is lower than 200, the statistical results will likely contain more noise than actual information. The items that

are field-tested by being embedded on the operational test forms as non-scorable items are administered to 200-250 applicants before they are made operational on subsequent test forms. This number of candidates is sufficient to determine if an item is functioning as expected; however, this number is not sufficient to support meaningful statistical analysis for DIF detection. Furthermore, statistical analysis of DIF (conducted when sample sizes permit) as well as statistical monitoring of item performance, is not used as a sole method of determining whether an item contains bias; but, rather, as another piece of information to guide additional content review of the flagged items.

The issue of a low sample size also applies to field testing conducted on stand-alone forms. There are generally insufficient numbers of participants in stand-alone field tests to allow for meaningful statistical analysis for DIF detection; statistical monitoring of item performance is conducted instead. In addition, the sample of examinees participating in stand-alone field tests is unlikely to be representative of the candidate population due to the self-selected nature of participation in field testing and potentially low test-taking motivation of the examinees. Regardless of whether statistical detection of DIF is feasible, best measurement practices call for a thorough review of the items for accuracy of content, job-relatedness, and freedom from bias. All items on the operational ALST forms underwent both content and bias reviews by the New York State educators and educator preparation faculty (see Chapter 3 for the description of Item Review Conferences).

2. Constructed-response items. Statistical analyses were conducted to ensure that ALST constructed-response items are of comparable difficulty.

At least 35 field test participants responded to each set of constructed-response items. A four-point scoring rubric was applied to each of the three CRIs in each set by two trained scorers using a focused holistic scoring process, for a total of 8 maximum points per CRI. In the test design for the CRI section, two of the CRIs were weighted as contributing 25% each, while the remaining CRI was weighted 50%, so the maximum score on the CRI section is 8 + 8 + 16 = 32.

Consequently, for each CRI set, a total of 32 raw score points was possible for each field test participant. Mean total scores across the sets of constructed-response items administered ranged from 15.0 to 22.7.

When examining the field test statistics obtained for the CRIs, the following criteria were used:

- Score distribution contains a reasonable spread of the various score points
- The overall difficulty of the CRI is in the desired range (not too hard and not too easy)

In addition, qualitative analyses were conducted based on the following sources of information:

- participant comments
- scorer comments

- score discrepancies
- number of blank, unscorable, or low-scoring responses

The comparable difficulty of each item set on constructed-response item forms was examined using Tukey's Studentized Range (Honest Significant Difference) Test for Total CRI component score.

The results of the Tukey analysis define which groups of CRI sets are statistically equivalent in terms of the total component score.

If the results of the Tukey analysis placed the CRI sets in the same Tukey grouping, and if the qualitative reviews listed above identified no performance issues with the CRIs, then the CRI sets were considered to have the appropriate characteristics for inclusion in the final item bank and were available for blueprinting on operational-test forms. Constructed-response items that did not meet those statistical and qualitative criteria were deleted from the item bank.

As a result of the first round of field testing on field-test forms, out of six CRI sets field-tested, three sets of CRIs (three extended-response and six focused-response items) were identified for inclusion on operational test forms and three CRI sets were identified for further field testing to collect additional data prior to consideration for future operational use.

As a result of the second round of field testing on field-test forms, out of 14 CRI sets field-tested, 12 sets of CRIs (12 extended-response and 24 focused-response items) were identified for inclusion on the operational-test form and two sets were deleted (two extended-response and four focused-response items). In summary, 15 out of 20 CRI sets were identified for inclusion on the operational-test forms after both rounds of field testing. The mean raw scores of the CRI sets that were included in the operational item bank ranged from 15.0 to 19.3.

Conduct Marker Establishment Meeting

The establishment of marker responses is an integral part of the process of preparing for operational scoring. Marker responses are examples of each of the score points on the scoring rubric. The marker establishment process helps set the criteria and standards for the scoring of candidate responses to constructed-response items. This is accomplished through the identification of a set of responses exemplifying each of the score points on the scoring rubric.

The ALST scoring rubric was developed by the Pearson testing experts in conjunction with NYSED. The development of the rubric was primarily informed by content and psychometric considerations (i.e., the use of performance characteristics that would enable reliable assessment of the range of academic literacy skills defined by the NY P-12 CCLS for ELA and Literacy) and established best practice regarding professionally developed examinations for educator certification, as specified in the Standards for Educational and Psychological Testing (AERA,

APA, & NCME, 1999). The use of the rubric by trained, experienced scorers in the first ALST field test was used to finalize its structure and content for subsequent operational scoring.

Marker responses help to ensure that scorers are applying the performance characteristics and rubric accurately, fairly, and consistently to examinee responses. NYSED, New York State educators from the Content Advisory Committees, and Pearson were involved in the process of reviewing candidate responses and the approved scoring rubric in order to identify a set of responses to be used as marker responses for the training of scorers.

Following field testing, Pearson invited members of the Academic Literacy Skills Test (ALST) Content Advisory Committee (CAC) to review field-test responses to the constructed-response items for the purpose of establishing marker responses for each item. The goal of each committee was to identify, for each constructed-response item, a set of four responses that clearly defined each of the four possible score points.

The establishment of marker responses involved the review of a number of responses across the entire range of the rubric and discussion of the characteristics of each response as they related to the performance characteristics that are included in the approved rubric. After being oriented to the performance characteristics and scoring rubric, committee members independently reviewed a series of responses. For each score point, committee members were asked to establish a response to serve as an exemplar of the score point description. Committee members then discussed their choices in order to reach consensus as a group on a final set of marker responses.

The Marker Establishment Meeting for ALST was held in Malta, New York, on July 8–9, 2013. Four members of the ALST CAC (generally representative of the expertise and characteristics of the larger group) participated on July 8, 2013, and five members participated on July 9, 2013.

Introduction, orientation, and training. NYSED program staff provided information regarding the background, purpose, and policies of the ALST. Pearson provided an introductory orientation including:

- a general overview, a review of the marker response establishment activity and the responsibilities of committee members;
- an explanation of the focused holistic scoring process;
- an introduction to the types of scoring tools that are used for the assignment of scores, including performance characteristics and scoring rubric; and
- an introduction to the various types and functions of marker response sets that are used to orient scorers to score operational responses after each test administration.

Following the general orientation, the CAC met as a committee and proceeded under the guidance of NYSED and Pearson facilitators.

Materials. Each CAC member registered upon arrival by completing a Sign-in Sheet and a Confidentiality Agreement. CAC members received the following materials:

1. Orientation Manual
2. Framework and Test Design
3. NYSED-approved Constructed-Response Performance Characteristics
4. NYSED-approved Constructed-Response Item Scoring Rubric
5. Sample responses to the Constructed-Response Items

Procedures. After an initial orientation session, CAC members met to conduct the marker establishment activities. They received further information from a Pearson facilitator regarding the review process for the responses and the rubric to be applied. Before beginning the review tasks, committee members familiarized themselves with the test framework on which the constructed-response items were based, the constructed-response items they previously approved, and the NYSED-approved performance characteristics and rubric. Please see Appendix Q for the ALST performance characteristics and rubric.

Marker establishment began with a review of sample responses for the constructed-response item that would appear on the first operational-test form. Committee members independently reviewed a number of responses as potential markers for each score point description of the scoring rubric. For each score point, committee members were asked to come to consensus regarding a response that represented each score point description. If needed, committee members modified and/or created responses to establish markers. The marker responses established for the constructed-response items that appear on the first operational-test form are called the historic anchor set. This set of responses is used at the beginning of the training process for each scorer training session.

Outcomes. As approved by NYSED, Pearson used the marker responses identified by the ALST Content Advisory Committee as part of the orientation at the scoring sessions following test administrations. All participating CAC members signed off on the ALST marker responses.

Assemble Test Forms

Test form construction. On September 23, 2013, the new NYSTCE Academic Literacy Skills Test (ALST) began to be administered to candidates. Each test form was constructed in accordance with the ALST design and blueprinting specifications (see Chapter 2). At the inception of the program, a "base form" or the first operational form was created. Afterwards, "equivalent forms" or subsequent forms were created. The process for assembling the base and equivalent forms is described below.

Base form creation. The ALST base form was used for the Standard Setting activities (see Establish Performance Standards section of this chapter) and remained active until additional equivalent forms were available for use. The ALST base form was created using 40 selected-response items (32 scorable and 8 non-scorable) and one set of three constructed-response items (2 focused-response and 1 extended-response items).The ALST base form consists of five 8-item clusters, four of which are scorable. Three scorable clusters are based on an informational passage. The fourth scorable cluster is based on a literary passage. The non-scorable cluster is either informational or literary. Item selection guidelines for the ALST base form are described in detail below.

Item selection guidelines for ALST base form. The major focus in item selection is on content coverage and meeting the test blueprint (e.g., both informational and literary passages need to be included on each test form). Other content considerations include avoiding placing items on a test form in which information in one item provides information regarding the correct answer for another item.

Statistical performance of the items collected during stand-alone field testing (see section on Field Testing in this chapter) was also taken into account when item selections were made. Specifically, SRI p-values and point-biserial correlations were considered. Items with high p-values (at or above .95), low p-values (at or below .30), or low point-biserial correlations (below .20) are typically not selected for use on operational test forms. Pertaining to the CRIs, descriptive information obtained from the stand-alone field-testing (including descriptive statistics, raters' and candidates' comments) were used to select CRIs included on the ALST base form. All content targets were reached for the ALST base form. For statistical targets, the overall form values also met the desired expectations.

Equivalent form creation. The ALST equivalent forms were created after a specified total number (cumulative) of examinees have taken the base form or other equivalent forms. New test form construction for the ALST is triggered when approximately two hundred candidates have received a test form. From that point onward, new forms are continuously built and published at a rate coinciding with the number of candidates taking the test during each program year. After the initial period, multiple test forms are typically simultaneously administered, with the order of selected-response items randomized for each examinee.

Item selection guidelines for ALST equivalent forms. Each new equivalent ALST form consists of 4 clusters from the base ALST form that will make up the scorable item set on the new test form. The nonscorable cluster for the new test form will be selected from previously unused clusters remaining in the item bank until no such clusters exist. At that point, the nonscorable cluster will be selected based on previous operational performance. Assuming acceptable performance in the nonscorable slot on the ALST base form, the nonscorable cluster will become scorable on the new test form. If the nonscorable cluster is an informational passage, then it will

replace one of the three scorable informational passage clusters. If the nonscorable cluster is a literary passage, then it will replace the scorable literary passage.

The same content requirements are required for the child ALST forms that are required of the base ALST form. In addition to the statistical guidelines for the base form, equivalent forms are constructed to have similar statistical profiles. That is, the average p-value of scorable SR items on the new form is similar to that of the average p-value of scorable SR items on the base form. This accomplishes continuity and consistency across different forms of the test.

ALST CRI sets are also designed to be comparable across test forms. Statistical and content considerations are used to establish the comparability of the CRI sets across forms.

All content targets were reached for the ALST equivalent forms. Likewise, all items met the desired expectations for statistical performance.

Establish Performance Standards

Standard Setting Panel

Function. The main function of the Standard Setting panel was to make an informed recommendation to the NYSED regarding what score on the ALST demonstrates the level of performance expected for just acceptably qualified candidates (i.e., hypothetical individuals who are just at the level of academic literacy skills specified by the Performance Level Descriptor).

Recruitment and Selection. The ALST Standard Setting Panel included practicing New York State public school educators who hold permanent or professional certification in New York State and New York State educator preparation faculty (including education faculty and arts and sciences faculty). Since the ALST is taken by candidates seeking certification in multiple fields, individuals certified and practicing in any of the fields for which the test is required were eligible to participate. Special emphasis was placed on identifying individuals who share the Regents' commitment to ensuring that there is an effective teacher in every classroom, who believe in the goal of college- and career-readiness for all students, who have a track record of promoting student achievement and growth, and who have a solid understanding of the New York State Common Core Learning Standards.

Composition. The ALST Standard Setting Panel is composed of 18 members including both certified and practicing New York State educators and educator preparation faculty. The Standard Setting panel included practicing and permanently certified public school teachers and college and university faculty, including those teaching undergraduate or graduate arts and sciences courses in which prospective educators were enrolled. Refer to Appendix K for a description of the characteristics of the Standard Setting panelists.

Overview. The purpose of the Standard Setting Conference is to determine what score on a test demonstrates a specified level of performance. Generally, the standard setting process begins

with a statement of the intended *performance level* – that is, a description of what people meeting the performance standard should know and be able to do. The goal is then to determine a *cut score* on an accompanying test that separates those who meet the performance standard from those who do not.

NYSED and Pearson convened the ALST Standard Setting Conference on November 12–13, 2013, at the Pearson office in Malta, New York. A total of 18 NYSED-selected and approved committee members participated in this meeting (nine certified and practicing New York State educators and nine educator preparation faculty) to recommend passing standards for the ALST.

Conduct Standard Setting Conference. A modified Angoff method and the extended-Angoff method were used to determine the ALST cut-score. After an orientation to these methods, the Standard-Setting Panel was led through three rounds of independent standard-setting ratings. Each panel member provided information about the expected performance of qualified educator candidates for the test items.

Standards for the following two performance levels were set by the standard-setting committee:

- Level I: the minimum threshold needed to pass the examinations for certification purposes. The Level I candidate has the minimum level of academic literacy skills a teacher needs in order to be competent in the classroom and positively contribute to student learning. The Level I candidate has partially mastered the academic literacy skills as defined by the Common Core Learning Standards for English Language Arts (please see Appendix R for a detailed description of this Performance Level Descriptor (PLD)).

- Level II: a rigorous, aspirational goal for program and candidates; a high benchmark to strive towards. The Level II Candidate has mastered the academic literacy skills necessary for effective teaching. The Level II candidate has mastered the academic literacy skills defined by the Common Core Learning Standards for English Language Arts (please see Appendix R for a detailed description of this Performance Level Descriptor (PLD)).

Materials. Each standard-setting committee member registered upon arrival by signing a Sign-in Sheet and a Confidentiality Agreement. Throughout the course of the conference, the standard-setting members received materials including the following (rating forms were duplicated for the two Levels):

1. Training Manual
2. Test Design and Framework
3. Performance Level Descriptors (PLDs) for Level I and Level II Candidates
4. Supplemental Materials for CRIs
5. Performance Characteristics and Scoring Rubric for Focused Response items (Assignments 1 and 2)

6. Marker Papers for Focused-Response 1
7. Marker Papers for Focused-Response 2
8. Performance Characteristics and Scoring Rubric for Extended Response
   Marker Papers for Extended-Response 3
9. Round 1 Item Rating Form: Selected-Response Items (SRIs)
10. Round 1 Item Rating Form: Constructed-Response Items (CRIs)
11. Item Difficulty Data Report
12. Round 1 Item Rating Summary: SRIs
13. Round 1 Item Rating Summary: CRIs
14. Round 2 Item Rating Form: SRIs
15. Round 2 Item Rating Form: CRIs
16. Item-Based Passing Score Summary Sheet: SRI
17. Item-Based Passing Score Summary Sheet: CRI
18. Test-based Judgment Form
19. Conference Evaluation Form

Procedures. There were five parts to the Standard-Setting process:

Part One: Simulated-Examination Session

Orientation session. Training was conducted to familiarize panelists with the materials and procedures for the simulated-examination session. They were given ample opportunity to ask questions and obtain clarification about the materials and procedures.

Simulated-examination session. For this activity, the panelists were asked to answer the selected-response items on the first ALST operational-test form without knowledge of the correct responses and to review and consider how they would respond to constructed-response assignments. They used an answer key to check their responses to the selected-response section of the test. The procedure for the simulated examination session was communicated to the panel members as follows:

1. Review the Test Design and Framework to understand the knowledge and skills that are measured on the test.

2. "Take" the test. Circle your answer to each scorable selected-response item in the booklet. After you complete all selected-response items, ask the facilitator for the Answer Key. When you get to the constructed-response assignments at the end of the Test Booklet, consider how you would respond to each item.

3. Use the Answer Key to check your answers to the selected-response items. Your responses are for your information only. You will not be scored by anyone else. All Test Booklets will be destroyed after the conference.

Part Two: Performance Level Descriptors (PLD) for Level I and Level II candidates

Orientation session. Part Two of the conference focused on reviewing Performance Level Descriptors (PLD) for Level I and Level II Candidates (please refer to Appendix R), and associated knowledge, skills, and abilities, as well as definitions of Just Acceptably Qualified Candidates (JAQC) for each respective Level. Training was conducted to familiarize panelists with the concepts of PLDs and JAQC.

Defining Just Acceptably Qualified Candidate (JAQC). For this activity, panelists were asked to conceptualize the Just Acceptably Qualified Candidate (JAQC) as defined by the Performance Level Descriptors (PLDs) for Level I and Level II Candidates (see Appendix R).

The following definitions of the JAQC for each Level were used:

Just Acceptably Qualified Candidate: Level I

A hypothetical individual who is just at the minimum level of academic literacy skills a teacher needs in order to be competent in the classroom and positively contribute to student learning.

Just Acceptably Qualified Candidate: Level II

A hypothetical individual who has mastered the academic literacy skills necessary for effective teaching.

Parts Three through Five (Rounds 1 through 3) were first conducted for Level I Candidates and then repeated for Level II Candidates.

Part Three: Round 1: Item-based Judgments

Item-based Judgments for Selected-Response Items

Orientation session. Training was conducted to familiarize panelists with the materials and procedures for making item-based judgments for selected-response items. They completed an item-judgment practice session, were given ample opportunity to ask questions, and obtained clarification about the materials and procedures.

Item-based judgments for selected-response items. Participants were asked to review each selected-response item in the Test Booklet and provide a judgment for each item regarding the expected performance of "just acceptably qualified candidates" at the appropriate Level. For each selected-response item they provided a judgment regarding the percent of a hypothetical group of "just acceptably qualified candidates" who would answer the item correctly. When making their judgments, panelists answered the following questions for Level I and Level II Candidates, respectively:

When making item-level judgments about Level I Candidates:

> **Imagine a hypothetical group of individuals who are just at the minimum level of academic literacy skills a teacher needs in order to be competent in the classroom and positively contribute to student learning.**

What percent of this group would answer the item correctly?

When making item-level judgments about Level II Candidates:

> **Imagine a hypothetical group of individuals who have mastered the academic literacy skills necessary for effective learning.**

What percent of this group would answer the item correctly?

The panelists entered their judgments on scannable forms, which were then collected by the Pearson facilitator.

Item-based Judgments for Constructed-Response Assignments

Orientation session. Training was conducted to familiarize panelists with the materials and procedures involved in making item-based judgments for the constructed-response assignments. They were given ample opportunity to ask questions and obtain clarification about the materials and procedures.

Item-based judgments for constructed-response assignments. The panelists were asked to review the CRIs, performance characteristics and scoring rubric for each CRI (see Appendix Q), and the marker responses for each CRI. Afterwards, they provided, for each CRI, a judgment regarding the total score (from two to eight, representing the sum of scores from two individual scorers applying a 4-point score scale) that would be achieved by a "just acceptably qualified candidate" in the appropriate Level. When making their judgments, panelists answered the following questions for Level I and Level II Candidates, respectively:

When making item-level judgments about **Level I Candidates:**

> **Imagine a hypothetical individual who is just at the minimum level of academic literacy skills a teacher needs in order to be competent in the classroom and positively contribute to student learning.**

What score represents the level of response that would be achieved by this individual?

When making item-level judgments about **Level II Candidates:**

> **Imagine a hypothetical individual who has mastered the academic literacy skills necessary for effective teaching.**

What score represents the level of response that would be achieved by this individual?

The range of possible scores is 2 to 8 (two scores are combined)

The panelists entered their judgments on the scannable forms, which were then collected by the Pearson facilitator. For a detailed explanation of the focused holistic scoring approach used to score ALST CRIs, please see the Scoring Section in Chapter 4.

Round 2: Item-based Judgments

Review and Revision of Item-based Judgments

Orientation session. Training was provided with an orientation to the materials and procedures for making Round 2 item-based judgments for selected-response items. The panelists were given ample opportunity to ask questions and obtain clarification about the materials and procedures.

Item-based judgments. The panelists were asked to review the Round 1 Item Rating Summary and the Item Difficulty Data Report (with p-values or percent of candidates answering the item correctly for the SRIs and descriptive statistics for the CRIs) from the first set of examinees who had taken the ALST. This additional information was offered in order to inform the panelists about item difficulty levels for a typical applicant. Panelists were cautioned about taking the information as indicative of performance of JAQC.

The Round 1 Item Rating Summary provided several types of information:

- A listing of the judgments provided by each panel member, by sequence number, to each selected-response item and constructed-response item.
- The median rating for each item and the distribution of item ratings made during Round 1 by all panel members.

The Item Difficulty Data Report provided for each item the percent of the candidates who answered the item correctly during recent administrations of the exam.

The following considerations regarding the Item Difficulty Data Report were also communicated to panel members:

- The candidates for whom results are presented in this document may not reflect the same proportion of all the types and capabilities of candidates in the population who will take the tests in the future.
- Panelist ratings need not necessarily match the p-value (percent who answered the item correctly) for the item obtained during the test administration. The data are useful for assessing the relative difficulty of the items.

- Panelists are encouraged to rely on their professional judgment in determining their ratings. The percentage of examinees who answered the item correctly is only one piece of information to help inform your judgments.

The panelists received a new set of rating forms on which to record any revisions they made to their Round 1 item-based performance level judgments, using the same rating procedures as in Round 1. The panelists were instructed to indicate ratings only for those items for which they were making revisions to their Round 1 ratings. The panelists entered their final item-level judgments on the scannable forms, which were then collected by the Pearson facilitator.

Part Five: Round 3: Test-based Passing Score Judgments

Orientation session. Training was conducted to familiarize the panelists with the materials and procedures involved in providing component-based judgments for the selected-response and constructed-response components of the test. They were then given ample opportunity to ask questions and obtain clarification about the materials and procedures.

Component-based judgments. The panelists were asked to provide a judgment regarding the expected performance of a just acceptably qualified candidate in the appropriate Level, on each component of the overall test (i.e., selected-response and constructed-response components) (refer to Appendix S for a sample judgment form).

To help panelists make their judgments, they were given information compiled from their item-based judgments and those of others on the panel. Specifically, item-based passing scores for the SRIs and the CRIs based on the cumulative ratings of all panelists were made available to the panel at this point on the process. In addition, panelists were informed what the passing rate would be based on these standards in the sample of the first set of examinees who took the operational ALST form.

Importantly, at this step the panelists were reminded to consider the purpose of the program, including the goals of the Board of Regents and the New York State Education Department, the information they had reviewed regarding the test and the test results, and the requirements for an educator receiving a certificate in New York State, and to use all of this information when making their test-based judgments.

Panelists' evaluation of process and evaluation results. Following all activities, panel members completed an evaluation form on the Standard Setting Conference. On a five point scale, panel members rated how well they understood the training, how confident they were in their judgments, how satisfied they were with the time provided to complete the work, how satisfied they were with the coordination and logistics of the meeting, and how satisfied they were with the performance Standard Setting process. Panel members were also provided space to make any additional comments regarding the Standard Setting Conference proceedings. The evaluations of all 18 panelists were positive: 16 out of 18 panelists gave a rating of "Well" or "Very Well" (4 or

5 on the scale from "1 – Not Well" to "5 – Very Well") when evaluating their understanding of the process, confidence in the results, and satisfaction in the coordination and logistics of the conference. Moreover, the comments were constructive: for example, one member noted that it is "great to include non-ELA/Literacy folks – this is an important perspective" and "in future planning/documentation, note exact bldg. location". Please refer to Appendix T for a sample evaluation form and ALST Standard Setting Evaluation Results.

Outcomes. Following the meeting, Pearson calculated the recommended cut-scores based on the ratings provided by the Standard Setting Panel members' final component-based ratings (i.e., Round Three) for both Level I and Level II.

In addition to providing panel-based passing scores, Pass Rate Analysis reports were provided to the New York State Education Department and the Commissioner of Education for considering the possible implications of using the passing score recommended by the expert panelists at the Standard-Setting Conference. These reports contain passing rates that reflect a compensatory scoring model in which higher scores on one section of the test (selected-response or constructed-response) may compensate for lower scores on the other section of the test. The full list of the Pass Rate Analysis reports and the list of elements contained within are provided below.

Pearson provided NYSED the following reports after the Standard-Setting conference:

- Pass Rate Analyses – Level I
- Pass Rate Analyses by Reporting Group – Level I
- Pass Rate Analyses – Level II
- Pass Rate Analyses by Reporting Group –Level II
- Interpretive Notes for Pass Rate Analyses

The reports listed above contained the following elements:

- the panel-based passing score for SRI and CRI sections and its associated standard error of measurement (S.E.M.) for Level I and Level II
    - The S.E.M. was provided for NYSED consideration of expected variation of examinee observed scores around their true scores.
- pass rate (percent of examinees reaching the cut-score) at or above the panel-based passing score, and plus and minus 2 S.E.M. for both Level I and Level II passing scores, based on the first administration of the ALST
- supplemental pass rate analyses by demographic subgroups (gender and race/ethnicity)
- supplemental pass rate analyses by New York State sectors (i.e., CUNY, SUNY, Private, Other)

NYSED Establishment of Performance Standards. The State Education Commissioner and NYSED considered the standards recommended by the Standard-Setting Committee and made the final determinations for Level I and Level II passing scores. The ALST Passing Score Panel's recommended performance standards for both components of the test at both specified performance levels were implemented by NYSED. The "cut-score" (Level I) serves as the minimum threshold needed to pass the examinations for certification purposes; it was set at a scaled score of 520. The "mastery cut-score" (Level II) is not used for determining whether a candidate has passed the certification examination; rather, it provides a rigorous, aspirational goal for programs and candidates – a high benchmark to strive towards; it was set at a scaled score of 564 for the first operational test form. Please refer to Commissioner's Memo from December 20, 2013 (Appendix U).

Chapter 4: Technical Properties of the ALST Scores

Evidence to support content-based, scoring-based, and generalization-based inferences with the ALST scores was gathered to document the extent to which test scores reflect the quality of test taker performance on the tasks delineated in the content domain (Kane, 2006). This chapter is organized in three parts. The first part describes steps taken to support content-based validity evidence (content representation and relevance of the content areas) and summarizes relevant sources of evidence. The other two parts include descriptions of processes to collect evidence to support test score inferences (scoring rules and procedures to derive scores for the intended uses) and their generalization (inference from observed scores to universe scores – true scores – of which measurement error is one indicator), respectively. More specifically, the second part describes the procedures used for scoring responses to the ALST selected-response items and constructed-response items, summarizes the development of the ALST reporting scale, and describes the approaches followed to maintain the ALST reporting scale. The third part summarizes evidence on ALST test score reliability and the consistency of Pass/Did Not Pass decisions to support the generalization of inferences with ALST test scores.

The following pieces of evidence support the use of the ALST scores for their intended uses and interpretations.

- Evidence supports that ALST items measure reading and writing academic literacy skills delineated on the test domain (i.e., the New York State P-12 Common Core Learning Standards for Language Arts) that a teacher needs in order to be competent in the classroom and positively contribute to student learning.

- Evidence supports that rules, guidance, and procedures for scoring candidates' responses to ALST items are appropriately described and applied to compute test scores. Rules for scoring responses are consistently and accurately applied. Scaling and equating procedures are clearly described and appropriately followed to report ALST scaled scores and maintain the ALST reporting scale.

- Evidence supports that ALST test scores are dependable measures of candidates' true scores. Also, evidence supports that ALST total scaled scores used to make minimum competency[16] decisions classify candidates with a psychometrically appropriate level of precision.

The remaining portions of the chapter describe the processes used to collect evidence, the pieces of collected evidence, and ways to interpret collected evidence for each of the three inferences (content-based, scoring-based, and generalization-based inferences).

[16] Minimum competency is denoted on Individual Score Reports as the "Pass" status.

Content-Based Validity Evidence

According to The Standards for Educational and Psychological Testing, validation of scores from credentialing tests such as the ALST "depends mainly on content-related evidence" (AERA, APA, & NCME, 1999, p. 157). Such content-based validity evidence is a logical process that establishes the connections between test content and job-related tasks. Content-based validity evidence for the ALST was gathered to ensure that the test content outlined in the ALST Framework is both representative and relevant to the content domain (AERA, APA, & NCME, 1999) as well as linked to the critical job-related tasks. Careful steps and procedures as summarized in Table 6 were taken throughout the test development process to ensure a strong correspondence between the content covered by the ALST and the tasks that are important to the job of an educator in the New York State. These steps and procedures are outlined in the paragraphs that follow.

The ALST test development process started with the development of the ALST Framework. As stated in the Standards for Educational and Psychological Testing, "the delineation of the test framework can be guided by theory or an analysis of the content domain or job requirements as in the case of many licensing and employment tests" (AERA, APA, & NCME, 1999, p. 37). As such, careful steps were taken in the early stages of the ALST Framework development to ensure that the competencies outlined in the framework are appropriate targets of measurement for the ALST. As a first step, a careful analysis of the job-related reading and writing skills implicit in the New York State Teaching Standards (Appendix A) was conducted (see Chapter 1 for the detailed description of the multiple surveys used in the development of these Standards). As a second step, the reading and writing skills expected of New York State students (and thus, teachers) – as operationalized in the New York State P-12 Common Core Learning Standards for English Language Arts and Literacy (NY P-12 CCLS for ELA and Literacy; Appendix L) – were carefully analyzed by NYSED and Pearson. Third, the ALST Framework (Appendix I) was drafted based on these analyses of the job-related tasks of an educator in New York State (see Chapters 2 and 3 for the detailed description of the ALST Framework). Fourth, correspondence was established between the contents of the ALST Framework and job-related reading and writing skills operationalized in the NY P-12 CCLS for ELA and Literacy. This correspondence was formally documented in the Content Correlation Table (Appendix N) and reviewed by New York State educators and educator preparation faculty. As a fifth and final step in the framework development process, the draft ALST Framework was reviewed for relevance and job-relatedness by the New York State Education Department (NYSED)-designated experts (Appendix J).

After the ALST Framework had undergone initial development, further steps were taken in order to gather additional content-based validity evidence. As a first step in this process, the ALST Framework was reviewed by the New York State educators and educator preparation faculty (Appendix K) during the Framework Review Conference (see the detailed description of the conferences in Chapter 3). During this effort, a Bias Review Committee (BRC) reviewed the

content of the ALST Test Framework and was charged with bias prevention, which is defined as (1) excluding language or content that might disadvantage or offend an examinee because of her or his gender, race, nationality, national origin, ethnicity, religion, age, sexual orientation, disability, or cultural, economic, or geographic background, and (2) including content that reflects the diversity of New York State. In addition, the BRC also reviewed the ALST Framework for appropriateness and job-relatedness. Following the Bias Review, a Content Advisory Committee (CAC) reviewed the framework for content appropriateness and job-relatedness, incorporated bias-related comments from the BRC, and made final recommendations to NYSED regarding the ALST Framework. As a second step, Content validation (CV) surveys were conducted with the New York State educators and faculty in order to gather additional evidence confirming that the content of the ALST Framework is appropriate and job-related. Verification of the on-the-job importance of the competencies and performance indicators comprising the ALST Framework adds a measure of convergent evidence of the validity of the inferences made from ALST scores. A full report containing CV Survey Results for the ALST can be found in Appendix O.

Throughout the development process, starting with the Teaching Standards (see Chapter 1), job-relatedness of the competencies measured by the ALST has been a reoccurring theme. The fourth and final step in the process of collecting content-based validity evidence for the ALST Framework – conducting the Job Analysis Study – was essential for ensuring that the tasks identified by the New York State educators as critical for performing a job of an educator were covered in the ALST Framework, thereby rendering ALST competencies and performance indicators as appropriate targets of measurement for the ALST. Empirical evidence garnered through the multi-step Job Analysis Study supported the logical chain from critical job-related tasks performed by the NYS educators, to the knowledge, skills, abilities, and other characteristics (KSAOs) (referred to as "competencies" in the ALST Framework) required for performing those tasks, to the inclusion of these critical competencies in the ALST frameworks. The Job Analysis Study is described in detail in Chapter 3. A full report can be found in Appendix M.

After sufficient content-based validity evidence for the appropriateness and job-relatedness of the ALST Framework was collected, the process of developing items matched to the framework commenced. Analogous to the framework development and validation, several steps were taken to build content-based validity evidence for the items. First, Assessment Specifications (Appendix P) were created in accordance with the job-related competencies outlined in the ALST Framework and served as a foundation for creating prototype ALST items by NYSED and Pearson. Second, after a sufficient number of ALST items were created, they were taken to the Item Review Conference (see the detailed description of the conferences in Chapter 3). During this effort, a Bias Review Committee (BRC) reviewed the content of the ALST items and was charged with bias prevention, which is defined as (1) excluding language or content that might disadvantage or offend an examinee because of her or his gender, race, nationality, national

origin, ethnicity, religion, age, sexual orientation, disability, or cultural, economic, or geographic background, and (2) including content that reflects the diversity of New York State. In addition, the BRC reviewed ALST items for appropriateness and job-relatedness. Following the Bias Review, a Content Advisory Committee (CAC) reviewed the items for content appropriateness and job-relatedness, incorporated bias-related comments from the BRC, and made final recommendations to NYSED regarding revisions to the ALST items.

Careful steps and procedures taken throughout the test development process ensured a strong correspondence between the content delineated in the ALST Framework and the tasks related to the job of an educator in the New York State (NYS). In addition, this content-based validity evidence supports that ALST items measure reading and writing skills that an educator needs in order to be competent in the classroom and positively contribute to student learning. Other test development activities, such as field testing, marker establishment, and establishment of the performance standards, further contribute to the validity of the ALST test scores and support the notion that the ALST test scores can be used to ensure that the NYS educator candidates have the minimum knowledge, skills, and abilities to successfully teach their students literacy skills as established by the NYS Teaching Standards. For the detailed description of these test development activities, please see Chapter 3.

Table 6. Academic Literacy Skills Test (ALST) Content Validity Evidence and Linkage to Test Purpose

ALST Purpose: The main purpose of the Academic Literacy Skills Test (ALST) is to measure a teaching candidate's literacy skills (reading and writing skills) implicit in the New York State Teaching Standards and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning.			
Test Development Activity		**Chapter of this report**	**Supporting evidence**
Development of the ALST Framework	Analyze the job-related reading and writing skills implicit in the NYS Teaching Standards.	Chapter 1	Appendix A: The New York State Teaching Standards
	Analyze the job-related reading and writing skills as operationalized in the NY P-12 CCLS for ELA and Literacy.	Chapters 1, 2, 3	Appendix L: New York State P-12 Common Core Learning Standards for English Language Arts and Literacy (NY P-12 CCLS for ELA and Literacy)
	Draft the ALST Framework based on the job-related reading and writing skills identified through the analysis of the relevant policy materials.	Chapters 2, 3	Appendix I: Academic Literacy Skills Test (ALST) Test Design and Framework
	Establish correspondence between the ALST Framework and the NY P-12 CCLS for ELA and Literacy.	Chapter 3	Appendix N: ALST Content Correlation Table
	Conduct preliminary review of the draft ALST Framework by NYSED-designated experts.	Chapter 3	Appendix J: ALST Curriculum Specialists Designated by NYSED to Review Preliminary Frameworks
Validation of the ALST Framework	Conduct Framework Review Conferences with the involvement of the NYS educators and educator preparation faculty. - Bias Review Conference (charged with detecting bias) - Content Advisory Committee (charged with ensuring	Chapter 3	Appendix K: Characteristics of the Review Committees

Validation of the ALST Framework	that test content is appropriate and job-related)		
	Conduct Content Validation Survey to gather evidence that the content of the ALST Framework is appropriate and job-related	Chapter 3	Appendix O: ALST Content Validation Survey Results
	Conduct Job Analysis Study to gather empirical evidence of the logical chain from critical job-related tasks performed by the NYS educators, to the knowledge, skills, abilities, and other characteristics (KSAOs) (referred to as "competencies" in the ALST Framework) required for performing those tasks, to the inclusion of these critical competencies in the ALST frameworks.	Chapter 3	Appendix M: New York State Educator Job Analysis Report: Volumes I and II. Human Resources Research Organization (HumRRO) (2014)
Development of the ALST items	Develop Assessment Specifications in accordance with job-related competencies outlined in the ALST Framework and create items	Chapter 3	Appendix P: ALST Assessment Specifications
Gathering content validity evidence for the ALST items	Conduct Item Review Conferences with the involvement of the NYS educators and educator preparation faculty. - Bias Review Conference (charged with detecting bias) - Content Advisory Committee (charged with ensuring that test items are appropriate and job-related)	Chapter 3	Appendix K: Characteristics of the Review Committees

Scoring-Based Validity Evidence

Scoring-based validity evidence for the ALST was gathered to ensure that the test scores reflect the quality of test takers' performance on the tasks delineated in the content domain (Kane, 2006). The primary focus of scoring inferences is on the rules and procedures for developing the ALST scores that are to be used for interpretations and decisions. Careful steps and procedures, as summarized below, were taken to ensure valid inferences from test scores. The section is organized in four parts.

- Scoring: the first part describes the use of appropriate procedures for scoring responses for selected-response items and constructed-response items.
- Scaling: the second part summarizes the appropriate development of the reporting scale.
- Equating: the third part describes the approach followed to maintain the reporting scale.

Evidence supports that rules, guidance, and procedures for scoring candidates' responses to ALST items are appropriately described and applied to compute test scores. Rules for scoring responses are consistently and accurately applied. Scaling and equating procedures are clearly described and appropriately followed to report ALST scaled scores and maintain the ALST reporting scale. The following paragraphs introduce the several pieces of the process to collect evidence.

Scoring Process. The ALST reports scores for the selected-response component (i.e., *Reading*) and the constructed-response component (i.e., *Writing to Sources*). The scoring process is rigorously performed in accordance to professional standards (AEAR, APA, NCME, 1999). A description of the scoring process for each component is presented below.

Selected-response items. The responses to the selected-response items on each ALST form are recorded digitally via computer-based testing. The responses are electronically scored based on the established answer keys and raw scores are computed from the set of operational scorable items

Constructed-response items. The responses to the ALST constructed-response assignments are scored using a method known as focused holistic scoring. In this method, scorers judge the overall effectiveness of each response while focusing on a set of characteristics that have been defined as important for the test. These performance characteristics guide scorers in the assignment of holistic scores using uniform criteria for all responses. The performance characteristics are included in the test directions to examinees that accompany the assignment.

Performance characteristics and scoring rubrics were developed for each type of ALST constructed-response item through a peer review process using New York State educators (see

Appendix K). Performance characteristics and score rubrics were later reviewed by New York State educators and teacher educators in conjunction with test development committee meetings.

Candidate responses to each constructed-response are rated on the scoring rubrics developed for that item type. Within the range of scores (from "0" to "4"), a "0" represents a blank or unscorable response to an item, while a "1" is assigned to a response that reflects a lack of relevant skills (for focused-response items) or a lack of argumentative writing skills (for extended-response items). A score point "4" is assigned to a response that reflects a strong command of relevant skills (for focused-response items) or a strong command of argumentative writing skills (for extended-response items). Responses that receive a particular score point reflect a range of knowledge and skills across that score point. For example, among the most competent responses, there are some that represent a "high 4" (strong) as well as those that represent a "low 4" (clearly superior responses, but they are not quite as strong as the "high 4"). This range of ability holds true within each of the other points on the scoring scale.

Please refer to Appendix Q for the scoring rubrics (with their associated set of performance characteristics) for each type of ALST CRI.

The following characteristics guide the scoring of responses to the ALST *focused-response* assignments:

1. Content
2. Analysis, Evaluation, and Integration
3. Command of Evidence
4. Coherence and Clarity

The following performance characteristics guide the scoring of responses to the ALST *extended-response* assignment:

1. Content and Analysis
2. Command of Evidence
3. Coherence, Organization, and Style
4. Control of Conventions

Scorers. Holistic scoring sessions for ALST are conducted by trained scorers working under the supervision of a Chief Reader. The Chief Reader is responsible for the overall management of the scoring session. The Chief Reader leads scoring sessions, provides training to scorers, and evaluates and monitors scorer performance. In addition, the Chief Reader identifies and develops orientation materials and qualifies scorers who evaluate candidate responses.

Scorers work under the supervision of the Chief Reader. Candidate responses are evaluated by at least two qualified scorers (i.e., each paper is rated by at least two scorers trained and supervised by the Chief Reader; the Chief Reader also provides third reads as necessary – as described below).

The New York State Education Department has established the following qualifications for ALST scorers. Individuals are eligible to become ALST scorers if they:

1. have a permanent Professional or Initial New York State teaching certificate
   AND
2. are currently teaching or have taught in New York State schools (public or private) within the last three years (including regular substitute teaching)
   OR
   are or have been educators from colleges and universities involved in teacher preparation within the last three years
   OR
   are an experienced and currently qualified NYSTCE scorer.

Experienced and currently qualified scorers are eligible to continue serving as scorers only if they have participated successfully as a scorer within the past 12 months and have also participated in a pre-scoring training session designed to orient scorers to the Regents' Reform Agenda. The training module for all new and continuing scorers based on the NYSTCE and the Regents' Reform Agenda:

- emphasizes key linkages of the new NYSTCE assessments with the Regents' Reform Agenda
- provides assessment specific information relative to the New York P-12 Common Core Learning Standards and other reform efforts
- provides clear connections of the new teaching assessments with the holistic scoring process
- provides scorers with materials and resources aimed at reinforcing their understanding of those connections.

Scoring Procedures. Responses to the ALST constructed-response items are scored using an independent-scorer scoring model. In this model, each response is read and scored independently by at least two calibrated scorers. Each response is given a focused holistic score by the two scorers on a scale of "0" to "4," with "4" as the highest score on the scale. Thus, the range of possible combined scores from the scorers will be from "2" to "8," respectively. Scorers assign "U" to responses that are unrelated to the assigned topic, illegible/inaudible/incomprehensible, in a language other than the language required by the constructed-response item, not of sufficient length to score, or merely a repetition of the assignment. Scorers assign "B" when there is no response to the assignment.

After completing orientation and calibration, which consists of familiarizing scorers with the scoring procedure, orienting them to the scoring rubric, and ensuring that all scorers are oriented to scoring candidate responses consistently, fairly, and in accordance with the rubric, scorers are

authorized to begin operational scoring. Each candidate response is then scored independently by two scorers.

The pair of scores assigned to each response is then compared and any responses needing further consideration are identified. If the first and second scores are identical or differ by one point, the two scores are added to provide the candidate's raw score. If the two scores differ by two points, the response is reviewed by a third qualified scorer. If this third reader's score matches the score of one of the first two scorers, the two identical scores are added to provide the candidate's final score. If the first two scores differ by two points and the third reader's score is in between the first two scores, then the third scorer's score is doubled to provide the candidate's raw score (note that this is equivalent to taking the average of the three scores). If the first two scores differ by more than two points, the Chief Reader assigns a score, and that score is doubled to provide the candidate's raw score.

ALST Reporting Scale. Design and development of reporting scales take into account multiple pieces of information and considerations. The following description first introduces a framework to conceptualize the general aspects of the ALST scoring scale. Then it moves to provide detailed information about characteristics of reporting scales. The third part is devoted to presenting information on the function adopted to convert raw scores to scaled scores. The final section includes a description of how constructed-response items are handled within the framework of the reporting scale.

Figure 3 shows a schematic representation of general scoring aspects for the Academic Literacy Skills Test (ALST). ALST raw scores for selected-response (Reading: X_R) and constructed-response (Writing to Sources : X_W) components are converted to scaled scores separately. The scorable portion of the ALST is composed of 32 scorable selected-response items and 3 scorable constructed-response items – two are relatively short focused-response items, and one is an extended-response item. X_R is computed by scoring the item responses (1 if correct and 0 if incorrect) for the scorable selected-response items. X_W is computed by linearly weighting the scores on the constructed-response items based on their rubric scores (0 to 4)[17] from each of two scorers. For ALST, the two focused-response items are each worth 25% of the total Writing component score, while the extended-response item is worth 50% of the total Writing component score. To comply with this weighting model, the scores from the extended-response item are doubled, so the maximum number of points on the Writing section is 32 (8 points for each of the two focused responses, plus 16 points for the extended response).

17 Detailed information on the scoring procedure followed for the constructed-response items can be found in the previous Scoring Process section of this chapter.

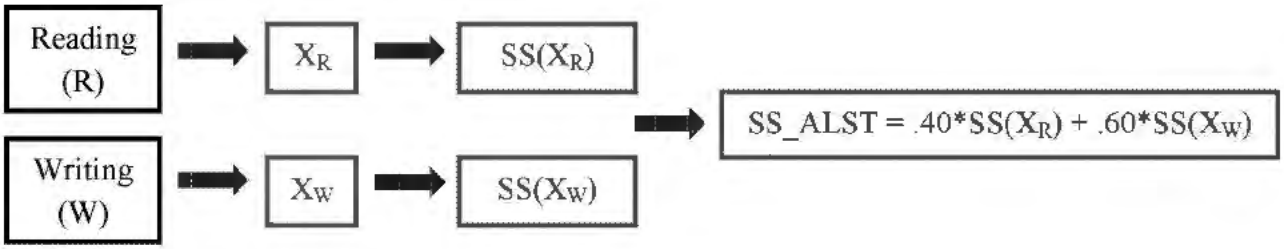

Figure 3. General aspects of ALST reporting scales

Figure 3 shows that raw scores are transformed to scaled scores separately for each component before they are combined (in a compensatory manner) to render the scaled total score. The transformation of the raw-to-scaled scores for the components is carried out with a scaling process that is described later in this section. The scaled scores from the selected-response ($SS(X_R)$) and constructed-response ($SS(X_W)$) components are then combined to compute the total scaled score (SS_ALST) following a defined weighting scheme: the selected-response component counts for 40% and the constructed-response component counts for 60% of the total score.

The goal of the ALST reporting scale is to provide to candidates their total test status (Pass/Did Not Pass) based on their total scaled score (SS_ALST). Table 7 shows a high-level summary of characteristics of the ALST reporting scale. The table presents two scales to better understand the scaling process for the ALST. The main scale is the scale used to report the total scaled score (SS_ALST) and it is the scale where the Pass/Did Not Pass decisions are made. The component scales are used as interim steps to compute the scaled scores for the selected-response component ($SS(X_R)$) and the constructed-response component ($SS(X_W)$), respectively (see Figure 3).

Table 7. Computation of the ALST Scale

Characteristic	Main Scale		Component Scale		
	Combined Reading & Writing		Reading		Writing
Use	Pass/Did Not Pass Decision		Component computation (weight = 40%)		Component computation (weight = 60%)
Range	400 to 600 points		400 to 600 points		400 to 600 points
Minimum Observable Scaled Score	Truncated at 400 points		Truncated at 400 points		Truncated at 400 points
Maximum Observable Scaled Score	Fixed at 600 points for perfect raw score		Fixed at 600 points for perfect raw score		Fixed at 600 points for perfect raw score
Scaled Cut Score	520 points		520 points		520 points

The reporting scales were developed to a standard set of specifications so it is easy for all stakeholders (candidates, institution representatives, and NYSED) to understand the meaning of a score. Table 7 shows that the total scaled score is reported on a scale that ranges between 400 and 600 points, with the passing cut score set at 520 points. The highest scaled score (600 points) has a one-to-one correspondence with the perfect test raw score. This upper bound is set by fixing the perfect test raw score to correspond to a score of 600 scale score points. The lowest scaled score (400 points) does not have a one-to-one correspondence with a raw score of zero points, so it can correspond to any raw score point that transforms to a scaled score of 400 or less using the linear equation described in the next section. This truncation keeps the reported scaled scores from taking values outside the range defined for the scale (Kolen & Brennan, 2004). Note that there are very few candidates who obtain scores that would be below 400 if not truncated.

A very important characteristic of the ALST scale is that it specifies a point (with meaning developed through the standard setting process) that identifies those teaching candidates who possess literacy skills (reading and writing skills) implicit in the New York State Teaching Standards and reflecting the minimum knowledge, skills, and abilities an educator needs in order to be competent in the classroom and positively contribute to student learning.

The scaled cut score of 520 is the cut for making the Pass/Did Not Pass decision[18], and this definition applies across not only the ALST, but all other NYSTCE assessments. A common scale in a fixed range applied to all new and redeveloped NYSTCE assessments achieves consistency in reporting and score interpretations. A fixed scaled cut-score of 520 allows for comparability across different test forms and different tests.

Developing the Scaling Function. The ALST scaled scores are computed using a general linear scaled score model. Previously, NYSTCE total scores were reported on a 100-to-300 scale, with 220 as the passing score. ALST and other tests will now be reported on the 400-to-600 scale, with 520 as the passing score.

Equation 1 shows a rendition of the general linear scaled score model applied to the ALST.

$$ScaledScore = \beta_1(RawScore - RawScoreCut) + 520 \quad \text{(Equation 1)}$$

The multiplicative constant (β_1) depicts the rate of change on which scaled scores vary relative to their corresponding raw scores. For the ALST the multiplicative constant can be computed with Equation 2.

$$\beta_1 = \frac{600-520}{MaxRawScore-RawScoreCut} \quad \text{(Equation 2)}$$

In Equation 2, the numerator computes the distance (in scaled score units) between the maximum scaled score and the scaled cut score. These two values are 600 and 520, respectively. Analogously, the denominator computes the distance on the raw score metric.

The above two equations can be applied to each component of the test to compute the scaled score for each component based on their raw scores. The results are then weighted appropriately to compute the scaled scores for the total test (see Figure 3).

For criterion-referenced tests with reporting scales based on performance standards anchored to a particular value, the setting of an anchor point on the scale is typically carried out by first assigning the raw cut score resulting from a standard setting procedure to correspond to a particular scaled score, such as 520 points. This anchoring typically is carried out with the base form and no equating is necessary. Once the reporting scale has been developed, it is customary to allow for administrations of new alternate test forms. An equating procedure is necessary to find the new raw cut score that is equivalent to the initial raw cut score (from the base form) to control for any differences in test form difficulty from form to form.

18 Modified-Angoff and extended-Angoff methodology were followed to determine the ALST cut score. See section on Establishing Performance Standards in Chapter 3.

Maintaining the ALST Reporting Scale for the Selected-Response Component. Once test users have become familiarized with the meaning of particular scaled test scores and organizations have adopted such scores for their decision-making process, changes to the reporting scale would become consequential to the inferences and use of the test scores. Test equating is a widely used approach to maintaining reporting scales for educational and licensure and certification testing programs (Kolen & Brennan, 2004). Testing programs often develop new forms along their lives and use equating methodologies to reports test-takers' results on their established scales.

ALST maintains a consistent reporting scale using test equating methods based on classical test theory. Various alternate forms are necessary to limit item exposure due to the large number of examinees taking the ALST and the number of examinees that retake the ALST. ALST alternate forms are developed to meet the same test blueprint specifications as the original test form, while maintaining, to the extent possible, the equivalent test difficulty (see Chapter 3 for more information on forms construction). Equating is the statistical procedure used to place alternate test forms on a common score scale (AERA, APA, & NCME, 1999). The purpose of test form equating is to compensate for differences in test difficulty and ensure that test scores from different test forms, after equating, have the same meaning. Therefore, differences in test scores are a reflection of differences in the ability level of the examinees and not of differences in test form difficulty. This is an important aspect of the technical quality of the test because, despite best efforts to create parallel test forms composed of items with comparable item statistics, it is not always possible. Equating allows scores from different test forms to be used interchangeably.

Kolen & Brennan (2004) list three data collection designs for test equating: single group design, random group design, and common-item non-equivalent groups design. The ALST uses the single group design to collect data for carrying out the equating of the raw scores for the selected-response component. The data collection design of the ALST allows for a very strong equating link between successive test forms because new alternate forms are constructed by embedding the items on each alternate form within the set of scorable and non-scorable items on the previously administered form. This approach permits the administered items to be scored in two ways – once using the set of scorable items on the previous form (base form) and again using the set of items that would appear on the alternate form when it is administered operationally. Hereafter, these two scores are referred as Y and X, respectively. Because the same set of examinees contributes to both computations, the equivalency between the two forms can be established in a robust manner.

The ALST equating uses a linear function and response data solely from the group of examinees who took the base form (Y). Response data from the base form are used to compute both the means and standard deviations of the scorable items on Y and the scorable items on the alternate form (X). In linear equating, two scores are equivalent if they are the same number of standard deviation units above or below the mean for some group of candidates (Angoff, 1984). A linear equation is used to relate the scores from the two forms by setting standard deviation scores, or

z-scores, to be equal on the two test forms (Kolen & Brennan, 2004). A raw score X that corresponds to a particular raw score Y can be calculated with Equation 3.

$$\frac{X_i - \bar{X}}{s_x} = \frac{Y_i - \bar{Y}}{s_y} \qquad \text{(Equation 3)}$$

where

X_i is a given raw score on the new alternate form

Y_i is the raw score equivalent to X_i expressed in the raw score metric of the previous form (base form)

$\bar{X}$ is the mean of the distribution of X_i scores

s_x is the standard deviation of the distribution of X_i scores

$\bar{Y}$ is the mean of the distribution of Y_i scores

s_y is the standard deviation of the distribution of Y_i scores

Equation 3 can be used to find raw score equivalents between points of the distributions of raw scores X and Y. In licensing and certification the point that matters the most is the Pass/Did Not Pass score. Equating is carried out with that simple goal, to find the raw cut score on an alternate form that corresponds to the raw cut score on the previously administered test form such that the scale score cut anchored on the ALST reporting scale remains the same. This mapping of raw score cuts between the distribution of scores X and Y can be done after rearranging Equation 3 and defining X_i as the raw cut score on an alternate form and Y_i as the raw cut score on the previous (base) form. Equation 3a shows the resulting calculation.

$$X_i = \bar{X} + (s_x/s_y)(Y_i - \bar{Y}) \qquad \text{(Equation 3a)}$$

Maintaining the ALST Reporting Scale for Constructed-Response Component. Maintenance of the CR reporting scale is accomplished through careful selection of the constructed response items prior to administration. Following initial field testing, CRI sets are selected that have equivalent mean scores according to analyses of variance and a post-hoc Tukey HSD test, as described previously in Chapter 3. As additional CRIs are developed and approved by New York educators, further field testing, including a representative operational constructed-response item, is conducted. The selected representative operational item that is chosen will have an average score in the middle of the distribution of scores for operational items. After further ANOVA and Tukey HSD analyses, those field test items that are grouped with the representative operational item are considered to be comparable in difficulty and are used on future operational test forms.

Generalization-Based Validity Evidence

The primary focus of generalization inferences is on finding chains to connect observed scores to universe scores by quantifying amounts of measurement error. This is an important distinction from scoring inferences, where the primary focus is on scoring rules and processes to derive the ALST score used to make decisions and interpretations. This section summarizes evidence on ALST test score reliability and consistency of Pass/Did Not Pass decisions to support the generalization of inferences from ALST test scores.

Evidence supports that ALST test scores reflect universe scores, which are expected values over conceptual replications of the measurement process. Also, evidence supports ALST scores (in particular the total scaled score use to make Pass/Did Not Pass decisions) classifies candidates with a level of acceptable precision. The remaining section summarizes several avenues that have been followed to collect evidence and discusses interpretations.

The Standards for Educational and Psychological Testing refer to reliability as the consistency of test scores for a group of candidates across administrations (AERA, APA & NCME, 1999). Test scores can fluctuate because of measurement error. Higher amounts of measurement error manifest on lower estimates of test score reliability. Also, there are a number of possible reasons for reliability estimates to achieve lower values. Some factors that affect reliability include the number of candidates, the number of test items, and test content. Within certification testing, additional factors, such as self-selection of candidates by test administration date and the variability of the group tested, can also affect reliability.

Ideally, score fluctuations caused by differences in the test itself are minimized, so that changes of test scores over time may be attributed to candidate factors. Because these tests are utilized to make high-stakes decisions, several indicators of decision consistency (the degree to which the same Pass/Did Not Pass decisions would be made from separate test administrations) and measures that indicate score reliability (consistency of scores across repeated administrations) are calculated for this program. Thus, the statistics presented not only consider the reliability of the test scores, but also indicate the reliability of the decisions made using the test results.

Test Form Reliability. Many examinees each year take the ALST. Some do not pass the ALST on their first attempt and may retake the ALST. Given the high annual volume of examinees and the group of examinees that retake the ALST, there are several different ALST test forms developed. See Chapter 3 for additional information regarding the development of test forms.

Multiple forms help prevent over-exposure of the test items and aid in keeping the test items secure. Each form is developed according to the test blueprint to include the appropriate content and types of items. The ALST uses Classical Test Theory reliability statistics, which are sample dependent; therefore, reporting reliability statistics by test form preserves the characteristics of each sample of examinees responding to a particular test form.

This technical report provides statistical information regarding each ALST test form and reports on candidate performance from the inception of the ALST through scores reported on or before May 30, 2014. The Technical Report Statistics by Test Form provides statistical characteristics of selected-response and constructed-response items. The Constructed-Response Item Statistics report provides scorer agreement rates and an interrater reliability statistic. The Total Scaled Score distribution provides a graphical display of scaled total test scores across all forms, as well as the number and percent of examinees at or above each score interval (including the passing standard). Specific details of each report are outlined below.

Several measures are employed to assess the reliability of ALST scores. The Technical Report Statistics by ALST Test Form, included in Appendix V, includes the number of examinees (including retakers) that took a form, the mean score on the form, and several statistics that provide reliability estimates. The reliability estimates are discussed below.

Decision Consistency. Decision consistency refers to the degree to which the same decisions would be made if an examinee took alternate forms of a test. Because it is not feasible to administer two forms of the test to the same candidate, decision consistency must be estimated by considering candidate performance on the single test form that the examinee happened to take. Given that the ALST is used to make mastery classifications (another term for Pass/Did Not Pass decisions), the consistency of such decisions becomes a primary source of validity evidence to support the generalizability of inferences based on the test scores (Crocker & Algina, 1986). In other words, it answers the question of what would the Pass/Did Not Pass decision have been if an alternate form of the ALST had been administered?

The Breyer and Lewis (1994) estimate of decision consistency is used. This method calculates decision consistency using a split-halves method in which one test form is split into two shorter forms that each represents the overall content of the whole test form. For the ALST, the SRI component of each split-half form is linked to the full version of the form to determine the cut on each split-half selected-response component, which is used for scaling purposes to calculate the component scale score. For the constructed-response component, two independent raters provide scores for each constructed-response item. One of the two scores for each constructed-response item is randomly assigned to each split-half form. The results are then scaled using the linear scaling model. The scaled scores for each component are weighted and combined to yield a scaled total test score (See sections on setting and maintaining ALST reporting scale in this chapter). Decision consistency is calculated using the scaled total test score. This method allows the use of both selected-response items and constructed-response items in the estimation of decision consistency. Decision consistency is reported in the range of 0 to 1, with estimates close to 1 indicating more consistent or reliable decisions.

The decision consistency estimates for the ALST test forms range from 0.79 – 0.82. There is no gold standard available to compare the decision consistency estimates. However, it is customary to compare the estimates achieved in a testing program against estimates observed in other

testing programs when no gold standard is available. This practice, though, is challenged by the scant number of testing programs reporting and using decision consistency estimates. In an effort to locate testing programs reporting decision consistency, a review of the literature identified two reports which are used to appraise the estimates for the ALST. The American Registry of Radiologic Technologists (ARRT) reports decision consistency estimates for the 2011 year ranging from 0.67 – 0.96, depending on the method used to calculate the estimate. And the National Athletic Trainers' Association Board of Certification, Inc. (2000) reports decision consistency estimates ranging from 0.78 – 0.97 on the different sections of their certification exam. The relative comparison between the ALST and the above testing programs indicated that the ALST decision consistency estimates fall within the range of values observed in other certification testing programs.

The location of the passing score within the distribution of test scores can affect the statistical estimates of decision consistency. The location of the passing score can contribute to classification errors resulting in false-positive and false-negative classifications of candidates. (Candidates that pass the test when they are expected to fail receive false-positive results and candidates that fail the test when they are expected to pass receive false-negative results.) When the mean score is near the cut score, the opportunity for false-positive and false-negative results increases, and therefore decision consistency estimates are lower (Crocker & Algina, 1989). As seen in the results in Appendix V, the mean score on each form of the ALST is quite close (within one standard error) of the cut score for all forms.

Kuder-Richardson formula 20 (KR20). The Kuder-Richardson index of item homogeneity (KR20) is an overall test consistency (reliability) estimate based on a single test administration (Kuder & Richardson, 1937). It is generally applicable to tests composed of selected-response items, where items are scored dichotomously, and is an indicator of internal consistency. KR20 is reported in the range 0 to 1, with a higher number indicating a greater level of consistency (reliability). Homogeneity refers to the degree to which the items on the test are consistent with one another. For the ALST, KR20 is computed for the selected-response component only.

The Technical Report Statistics by ALST Test Form in Appendix V shows the selected-response component has KR20 values ranging from 0.65 – 0.70. Classical reliability estimates in this range are considered low relative to the longer tests typically administered to students, which commonly have 60 items or more and have reliabilities greater than 0.80; however, the KR20 is a lower-bound estimate of internal consistency (Allen & Yen, 1979). The ALST has only 32 scorable selected-response items; it is not practical to expect a KR20 estimates comparable to those observed in student testing with longer tests.

There are a number of known limitations of the KR20. The KR20 estimate is influenced by the test structure. Tests that include multiple items tied to the same stimuli are known to have diminished reliability estimates (Wainer & Thissen, 1996). The ALST is composed of clusters of eight items all tied to the same stimuli.

The variance in the observed scores of the sample of examinees may also adversely impact KR20 estimates. The estimates can be suppressed if scores across the full range of the score scale are not observed. The distribution of scores for the ALST currently clusters around the cut score, and there are few observations at the low and high ends of the score scale (see Appendix V). Lower total test score variance, which is used directly in the computation of KR20, leads to lower KR20 estimates. Appendix V also provides items statistics for scorable items on the operational test forms. The p-values of the scorable ALST items ranged from 0.15 to 0.93, with an average p-value of 0.62. The point biserial correlations for the scorable ALST items ranged from 0.02 to 0.49, with an average point biserial correlation of 0.27.

Generalizability coefficient (G). The Generalizability (G) coefficient is a measure of the percent of total score variance that is attributable to persons (i.e., factors within the candidate, such as subject matter knowledge). It reflects the proportion of variability in individuals' scores that is attributable to true score variability rather than to measurement error (Brennan, 2001). It is reported in the range 0 to 1, with a higher number indicating a greater level of generalizability. The G-coefficient is generally applicable to tests composed of constructed-response items. It gauges the degree to which the results from one test form of the constructed-response items are generalizable to other forms, or other test administrations.

A Generalizability coefficient is presented for the constructed-response component of each ALST test form in Appendix V. The Generalizability coefficients range from 0.55 – 0.65. Given that there are only three constructed-response items within the component, the Generalizability coefficients reported for the ALST are relatively high.

There are several aspects that need to be considered when interpreting the magnitude of the generalizability coefficients for the ALST. Range restriction affects the magnitude of the variance components and can restrict the Generalizability coefficient. Since the reliability coefficient is computed by dividing the true score variance by the error variance, an inflated error variance would reduce the magnitude of the generalizability coefficient. A final consideration is the effect that the number of items has on the estimation of the error component. Because the number of items is inversely related to the amount of error variance, tests with a smaller number of items will have greater measurement error and a smaller generalizability coefficient.

Stratified alpha. Stratified coefficient alpha is an estimate of total test reliability for a test containing a mixture of item types (e.g., selected-response and constructed-response) (Qualls, 1995). Each item type component of the test is treated as a subtest. Internal consistency estimates for the separate subtests are combined to compute stratified coefficient alpha. Stratified coefficient alpha is reported in the range 0 to 1, with a higher number indicating a greater level of consistency (reliability). This statistical estimate is deemed most appropriate for estimating total reliability of tests with both selected-response and extended-response items for the ALST because it takes into account differences in the contribution of each component to the total test score and variance of both selected-response and constructed-response items.

The stratified alpha estimates for the ALST test forms range from 0.67 – 0.75. The stratified alpha estimate is based on the reliability of the selected-response and constructed-response components. As such, previously discussed considerations of the test design, characteristics of the sample, and concerns over limited variability of component and total test scores must be taken into account when considering the stratified alpha estimate as well.

The ALST reliability estimates are lower than typically observed on large-scale student achievement tests. As previously stated, since test score reliability depends, in part, on test length, it is not rare finding that the reliability estimates from certification testing achieve lower magnitudes than those from large-scale achievement testing which relies on longer tests (e.g., 50 or more selected-response items and 4 or more extended response items). It is important to keep in mind that the component scores are not used to make high-stakes Pass/Did Not Pass decisions. The critical Pass/Did Not Pass decision is made based on the scaled total test score. The decision consistency estimates for the entire ALST test forms range from 0.79 – 0.82.

Standard error of measurement (SEM). The Standards for Educational and Psychological Testing define the standard error of measurement as the standard deviation of candidate scores obtained from repeated administrations of tests or parallel forms of the test (AERA, APA, & NCME, 1999). The SEM is an estimate of the amount of variance in a score that results from factors other than examinee ability. The SEM is a range of scores around an observed score in which the true score, a score free of error, is expected to fall. The SEM is directly and inversely related to reliability; thus the higher the reliability of a test form, the smaller the SEM.

The SEM for all forms of the ALST range from 7.0 – 7.1 points on the ALST reporting scale for the total scaled test score. The total scaled score ranges between 400 points and 600 points.

Reliability of Constructed Response Items. The ALST constructed-response items are scored by two independent scorers. The scorers are required to undergo extensive training and pass a scoring test to qualify to score ALST constructed-response items. The scorers also use scoring rubrics and performance characteristics to maximize scoring consistency across scorers. Quality control scoring procedures are implemented to monitor the accuracy and consistency of raters' performance during each scoring session. Scorers that do not perform as expected on their levels of accuracy and consistency may be required to undergo additional training and pass an additional scoring test or may be dismissed.

Two methods used to monitor scorer performance, thus the reliability of the constructed-response scores, are agreement rates and interrater reliability estimates. Scorer agreement is the degree of agreement between constructed-response scores assigned by independent scorers. Independent scorers are in agreement if the scores they award are either exact or adjacent. Percent exact agreement is the percent of scorable responses for which the first two scorers are in exact agreement. Percent adjacent agreement is the percent of scorable responses for which the first two scorers are within one point of each other. Scorers are considered discrepant if the scores differ by more than one point. The percent of cases in which the first two independent

scorers are in agreement is computed as a measure of scorer agreement (reliability). Interrater reliability is computed as the intraclass correlation between the first and second score assigned to each response, corrected using the Spearman-Brown formula.

Reliability information about scorer performance is based on item type. As stated previously, constructed-response items that remain in the item bank following field testing are considered equivalent and thus interchangeable. This permits grouping items by item type for the purpose of reporting scorer reliability, which maximizes the information available for calculating reliability statistics. The ALST constructed-response items type groupings are defined by the task required by each type of item. One item type requires evaluation of an author's position, another item type requires interpreting and using information presented graphically, and another item type requires an extended length response.

The Constructed-Response Items Statistics Report, Appendix V, shows that scorers' percent agreement ranges from 96% to 97% and the interrater reliability estimates range from 0.66 to 0.76. The percent agreement rates measure degree of consistency of scoring the constructed-response items, indicating that the scores are a reflection of examinee ability rather than some unique property of the scorer. The interrater reliability estimate shows a moderate to high correlation between scorers, consistent with the agreement rates. [19]

The mean and standard deviation for each type of item is presented in Table 8 below. The mean and standard deviation are based on valid scores of 2 to 8 only. All attempts on each item type are reported.

Table 8. Means and Standard Deviations of Constructed-Response Items by Item Type

Type 1	8,607	5.6	1.2
Type 2	8,562	5.3	1.6
Type 3	8,526	5.6	1.4

[19] Agreement rates and interrater reliability estimates for ALST constructed-response items are provided only for constructed-response items that were administered to at least 10 candidates during the program year.

Test scaled score distribution. The Total Scaled Score Distribution based on all attempts made by each examinee is included in Appendix V. For the ALST, results are reported on a scale ranging from 400 to 600. A scaled score of 520 represents the passing standard for each test. The total scaled score distribution includes an inclusive list of observed total test scaled scores, in intervals of five scale-score points; the number of scores observed within intervals of five scale-score points; the number of scores observed at or above each scale-score interval; the percent of scores observed within intervals of five scaled-score points; and the percent of scores observed at or above each scaled-score interval.

The distribution of total scaled scores is quite sample dependent. That is, the characteristics of the distribution may change depending on the group of examinees included. For this reason, total scaled-score summary statistics and pass rates should be considered from several perspectives. It is informative to consider performance based on the examinees' *initial* attempt, *best* attempt, and the performance of examinees on *all* attempts. Reporting on initial attempt involves isolating the first attempt for each examinee. Reporting based on best attempt involves using only the highest total scaled score achieved by each examinee. The best attempt score may be a passing or failing score. This perspective is more representative of final outcomes. Reporting based on all attempts involves using every attempt made by every examinee. All passing and/or failing scores achieved by an examinee are included in the calculations of the summary statistics and pass rates. This perspective provides overall performance information.

There can be significant differences in the total scaled score summary statistics and pass rates depending on the use of initial attempt, best attempt, or all attempts. The inclusion of information for retake attempts suppresses the summary statistics and pass rates because the examinees included in the retake sample tend to have lower ability levels. The summary statistics and pass rates, reported by initial attempt, best attempt, and all attempts are contained in Table 9 below:

Table 9. Total Scaled Score Information for Initial Attempt, Best Attempt, and All Attempts

	Candidates		Deviation		
Initial Attempt	7,420	522.1	30.56	525.0	56.7
Best Attempt	7,420	524.9	29.42	528.0	63.5
All Attempts	8,622	520.9	29.88	523.0	54.7

As Table 9 demonstrates, the ALST mean total scaled score across all test forms, based on the best attempt of each examinee is 524.9, approximately three points higher than the mean score achieved on initial attempts and four points higher than the mean on all attempts. The median score on best attempt differs by three and five points on initial attempt and all attempts, respectively. The standard deviation is similar for score distributions based on best attempts and all attempts but slightly higher for initial attempts. The pass rates differ depending on attempt type as well: pass rates are lower by approximately 7% based on initial attempts and 9% based on all attempts than pass rates based on best attempts. The significant differences in the above results demonstrate the influence of the sample under consideration on ALST summary statistics and pass rates.

As mentioned above, examinees that retake the ALST generally possess lower ability levels. The scores achieved by retakers contribute variability to the distribution of scores, which has some benefits related to evaluating the reliability of the test form; however, these same scores work to suppress the performance of the entire sample. Information about retake performance is presented in Appendix V, Retake Summary Statistics and Pass Rates.

Chapter 5: ALST Score Reporting

After administration of the New York State Teacher Certification Examination (NYSTCE) Academic Literacy Skills Test (ALST), individual score reports (ISRs) are provided to candidates to inform them of their passing status and performance on the test. Candidates' test scores are also provided to the NYSED and, if applicable, to the institutions of higher education that the candidates indicate during exam registration. Aggregate score reports for all institutions are also provided to the NYSED, and institution-specific score reports are provided to score report contacts at New York State-approved educator preparation programs at the state's institutions of higher education.

Candidate Score Reports

All candidates that register online for the ALST may request that a score report be emailed to the address provided during the registration process on the score report date published on the NYSTCE website in Test Dates. In addition, the score report is available to the candidates on MyAccount on the NYSTCE website within 30 days of the testing date. Score reports are provided as PDF documents, which candidates may view, print, and save for their records. The score reports are available online for 45 days. After 45 days, candidates' Pass/Did Not Pass status is available to them through the testing history in MyAccount.

ALST Individual Score Reports (ISRs) include the following:

- The date the candidate took the test
- The candidate's address and the last 5 digits of their Social Security Number (SSN)
- The candidate's overall scaled score converted to a scale ranging from 400 to 600
- The scaled minimum passing score, which is 520
- The candidate's passing status based on the established passing standard
- Detailed performance information on each competency assessed by the ALST as a performance index ranging from 1 to 4

A sample ALST ISR can be found in Appendix W. The ALST ISRs are accompanied by an interpretive page to help candidates understand their reports. More information on how to understand the NYSTCE score reports is available to candidates here.

Score Reports for NYSED and Educator Preparation Programs

The New York State Education Department receives the following data after each test administration period:

Examinee Data File. This electronic data file provides score and registration information for each candidate who tested.

Update to ResultsAnalyzer™. The ResultsAnalyzer data file is updated with the release of score reports to candidates, preparation institutions, and the state.

In addition, The New York State Education Department receives the following data files annually:

Title II Program Completer Data File. This electronic data file provides program completer data generated as part of Title II pass rate reporting.

Statewide Results Data File. Electronic data files providing pass rate information by institution and sector for the following groups:

- All examinees
- Examinees in Title II program completer cohorts

Title II Reporting

Section 205 of Title II of the Higher Education Act (as amended in 2008) requires teacher preparation programs to report data on the assessments used for teacher certification or licensure by the state. These data include the number of test takers, the number who passed, the pass rate, the average scaled score, and the minimum passing score for each assessment. States must report these data for each IHE (traditional and alternative routes) and non-IHE-based alternative route. Each institution of higher education is required to provide data for candidates who are enrolled or have completed programs of professional teacher preparation. These data are then compared with testing records, and Assessment Pass Rate and Summary Pass Rate reports are generated for the purposes of federal reporting. As a test required for initial certification, data for the Academic Literacy Skills Test will be matched to candidate records and reported as part of annual Title II pass rate reporting.

References

Ahn. S., & Choi, J. (2004). Teachers' Subject Matter Knowledge as a Teacher Qualification: A Synthesis of the Quantitative Literature on Students' Mathematics Achievement. *Online Submission*. Available at http://eric.ed.gov/?id=ED490006.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Auguste, B. G., Kihn, P., & Miller, M. (2010). *Closing the talent gap: Attracting and retaining top-third graduates to careers in teaching: An international and market research-based perspective*. McKinsey. Available at http://www.bbp-action.org/wordpress/wp-content/uploads/2011/10/McKinsey-Report-Closing-the-Talent-Gap.pdf

Angoff, W. H. (1984). *Scales, Norms and Equivalent Scores*. Princeton, N J: Educational Testing Service.

Brennan, R. L. (2001) *Generalizability Theory*. New York, NY: Springer-Verlag.

Breyer, F. J., & Lewis, C. (1994). Pass-fail reliability for tests with cut scores: A simplified method (ETS Research Report No. 94–39). Princeton, NJ: Educational Testing Service.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 701-732). Westport, CT: American Council on Education and Praeger Publishers.

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational measurement: Issues and practices*, 31-44.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth Group.

Ehrenberg, R. G., & Brewer, D. J. (1995). Did Teachers' Verbal Ability and Race Matter in the 1960s? 'Coleman' Revisited. *Economics of Education Review*. *14*(1), 1-21.

Goldhaber, D., Perry, D. & Anthony, P. (2004). NBPTS Certification: Who applies and what factors are associated with success? *The Urban Institute, Education Policy Center.*

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, 84–117.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Human Resources Research Organization (HumRRO) (2014). *New York State Educator Job Analysis.*

Kane, M. T. (2006). Validation. In: R.L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport: American Council on Education/Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.). New York, NY: Springer Science and Business Media, LLC.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Qualls, L. A. (1995). Estimating the reliability of a test containing multiple item formats. Applied. *Measurement in Education 8*(2), 111–120.

Rice, J. K. (2003). *Teacher Quality: Understanding the effectiveness of teacher attributes.* Economic Policy Institute, 1660 L Street, NW, Suite 1200, Washington, DC 20035.

Society for Industrial and Organizational Psychology (SIOP). (2003). *Principles for the validation and use of personnel selection procedures: 4th edition.* Bowling Green, OH: Author.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,27*(4), 361-370.

Tchoshanov, M. A. (2011). Relationship between teacher knowledge of concepts and connections, teaching practice, and student achievement in middle grades mathematics. *Educational Studies in Mathematics, 76*(2), 141-164.

Uniform Guidelines on Employee Selection. (1978). *Federal Register*, 1978, 43, No. 166, 38290 – 38309.

Walsh, K., & Tracy, C. O. (2004). Increasing the odds: How good policies can yield better teachers. *National Council on Teacher Quality.*

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Whitehurst, G. J. (2002, March). Scientifically based research on teacher quality: Research on teacher preparation and professional development. *Paper presented at the White House Conference on Preparing Tomorrow's Teachers.*

Appendices

Appendix A: The New York State Teaching Standards

Appendix B: Structure of Pedagogy Content Standards Developed by Various Countries

Appendix C: Correlation of Sample National, State, and Local Teacher Standards with Commonly Addressed Topics

Appendix D: Invitation to comment on the preliminary draft of New York State Teaching Standards

Appendix E: Teaching Standards Survey Summary August 17th, 2010

Appendix F: Teaching Standards Survey Full Report August 17th, 2010

Appendix G: Model Core Teaching Standards: A Resource for State Dialogue

Appendix H: Teaching Standards Survey Results December 15th, 2010

Appendix I: Academic Literacy Skills Test (ALST) Test Design and Framework

Appendix J: ALST Curriculum Specialists Designated by New York State Education Department (NYSED) to Review Preliminary Frameworks

Appendix K: Characteristics of the Review Committees

Appendix L: New York State P-12 Common Core Learning Standards for English Language Arts and Literacy

Appendix M: New York State Educator Job Analysis Report: Volumes I and II. Human Resources Research Organization (HumRRO) (2014)

Appendix N: ALST Content Correlation Table

Appendix O: ALST Content Validation Survey Results

Appendix P: ALST Assessment Specifications

Appendix Q: ALST Scoring Rubric and Performance Characteristics

Appendix R: ALST Performance Level Descriptors (PLD): Levels I and II

Appendix S: ALST Standard Setting Test-Based Judgment Form (Round 3)

Appendix T: ALST Standard Setting Sample Evaluation Form and Results

Appendix U: New York State Board of Regents December 20th, 2013: Commissioner King Announces Scores Needed to Pass Teacher and Leader Certification Examinations

Appendix V: ALST Test Statistics

Appendix W: ALST Individual Score Reports (ISR)

Tables

Figures